UNIVERSITÉ DE GENÈVE
FACULTY OF SCIENCE
Department of Informatics

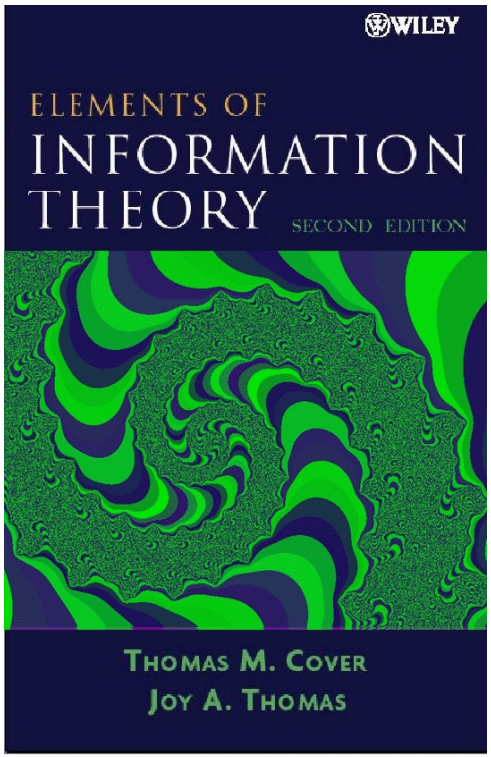
SIP stochastic information processing

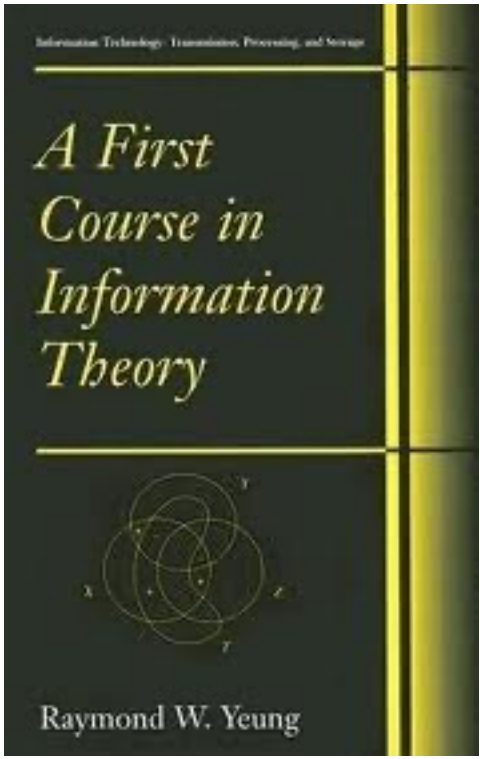
Information Theory
for Data Science and Machine Learning

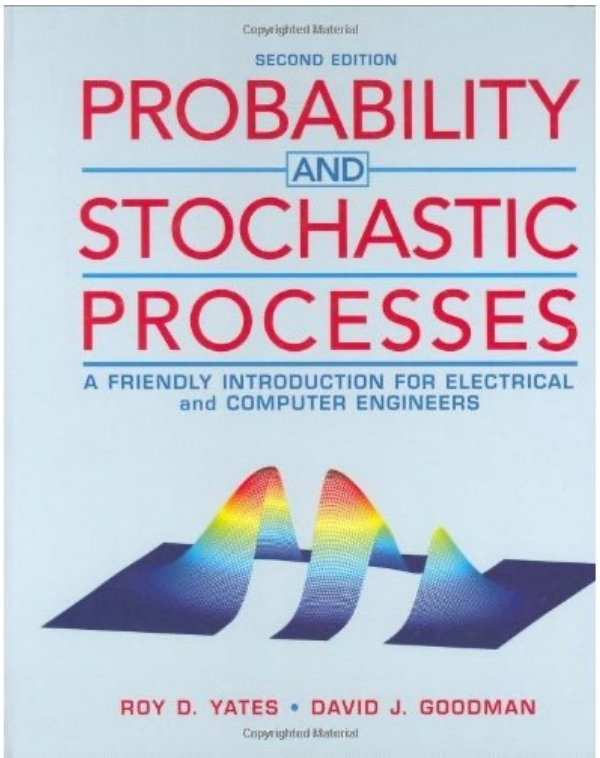
Slava Voloshynovskiy

http://sip.unige.ch/education/information-theory/

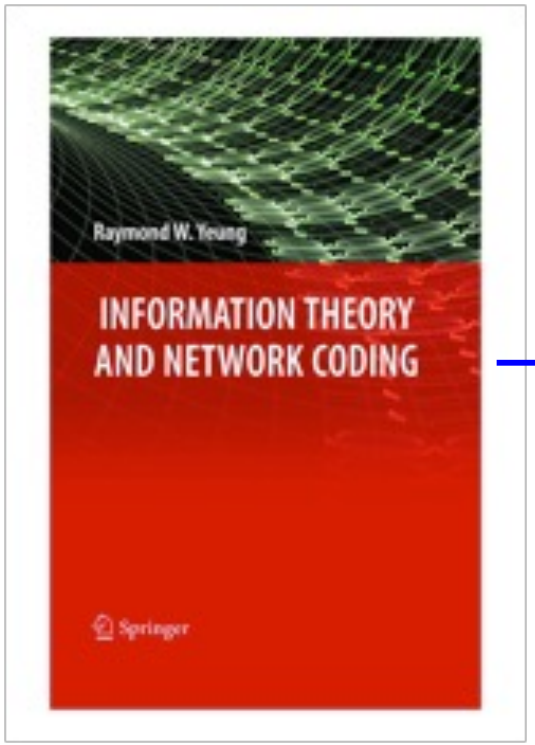
Recommended books







Recommended books



Free download as

Raymond W. Yeung

Information Theory and Network Coding

SPIN Springer's internal project number, if known

January 28, 2008

http://iest2.ie.cuhk.edu.hk/~whyeung/book2/

http://iest2.ie.cuhk.edu.hk/~whyeung/post/manuscript/main2.pdf

Copyright notice

This course uses in part some materials from previously mentioned recommended sources.

Content of course

- Theme 1 – Basic statistical data models
- Theme 2 – Information theoretic measures

Scope

- Information theoretic measures for discrete/continuous variables
  - Entropy
  - Conditional entropy
  - Joint entropy
  - Relative entropy (KL-divergence)
  - Cross entropy
  - Mutual information
  - Additional topics:
    - f-divergence
    - variational inference
    - practical computations from samples

Part I: discrete random variables and vectors

Part II: continuous random variables and vectors

Scope

Part I: discrete random variables and vectors

Scope

- Information theoretic measures for discrete/continuous variables
  - Entropy
  - Conditional entropy
  - Joint entropy
  - Relative entropy (KL-divergence)
  - Cross entropy
  - Mutual information
  - Additional topics:
    - f-divergence
    - variational inference
    - practical computations from samples

Entropy

Definition (entropy): *Entropy* of discrete r.v. $X \in \mathcal{X}$ with $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$

$$H(X) = -\sum_{x \in \mathcal{X}} p_X(x) \log_2 p_X(x) = -E_{p_X}\left[\log p_X(x)\right] \quad \text{(bits)}$$

$$H(X) = -\sum_{i=1}^{n} \Pr\{X = x_i\} \log_2 \Pr\{X = x_i\} \qquad E_{p_X}\left[g(X)\right] = \sum_{x \in \mathcal{X}} p_X(x) g(x)$$

- When the base of logarithm is α, we use the notation $H_\alpha(X)$
- Entropy measures the uncertainty or randomness of a discrete random variable
- The units of entropy

$$\text{bit} - \text{if } \alpha = 2$$

$$\text{nat} - \text{if } \alpha = e$$

- $H(X)$ depends only on the distribution of X but not on the actual value taken by X, hence also it is reflected as $H(p)$ or $H(p(x))$

Entropy

- Examples
  - Entropy of binary random variable with the equiprobable states

$$H(X) = -\left[\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right] = \log_2 2 = 1\ bit$$

  - Entropy of a random variable with 27 (a model of alphabet) equiprobable states

$$H(X) = -\frac{1}{27}\sum_{i=1}^{27}\log_2\frac{1}{27} = \log_2 27 = 4.7549\ bit$$

    - Note: in reality $H(X) = 3.98\ bit$

  - Entropy of a random variable with N equiprobable states/symbols

$$H(X) = -\frac{1}{N}\sum_{i=1}^{N}\log_2\frac{1}{N} = \log_2 N\ bit$$

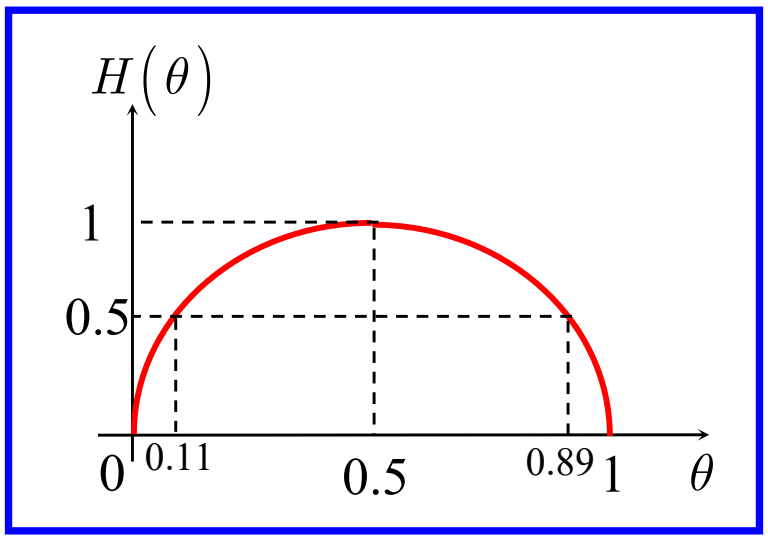
Entropie: binary random variable

- Binary random variable $X \in \mathcal{X} = \{0,1\}$

$$\Pr\{X = 0\} = 1 - \theta$$
$$\Pr\{X = 1\} = \theta$$

$$\Rightarrow \Pr[X = x] = \theta^x (1-\theta)^{1-x}$$

$$H(X) = -\theta \log_2 \theta - (1-\theta)\log_2(1-\theta) := H(\theta) \text{ or } H_2(\theta)$$



The function $H(\theta)$ is:

- Symmetric wrt $p = 0.5$
- Maximum $(H(p) = 1)$ at $\theta = 0.5$

The entropy is maximal, if the symbols are equiprobable.

Joint entropy

Definition (joint entropy): *Joint entropy* of two discrete random variables X and Y is defined by:

$$H\left(X,Y\right) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}\left(x,y\right)\log_2 p_{X,Y}\left(x,y\right) = -E_{p_{X,Y}}\left[\log p_{X,Y}\left(x,y\right)\right]$$

$$\mathcal{X} = \left\{x_1, x_2, \ldots, x_n\right\} \quad \mathcal{Y} = \left\{y_1, y_2, \ldots, y_m\right\}$$

$$H\left(X,Y\right) = H\left(Y,X\right)$$

Conditional entropy

Definition (conditional entropy): *Conditional entropy* of r.v. X given Y is defined by:

$$H\left(X\middle|Y\right)=\sum_{y\in\mathcal{Y}}p_Y\left(y\right)H\left(X\middle|Y=y\right)=-\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}p_{X,Y}\left(x,y\right)\log p_{X|Y}\left(x\middle|y\right)$$

$$\mathcal{X}=\left\{x_1,x_2,\ldots,x_n\right\}\quad\mathcal{Y}=\left\{y_1,y_2,\ldots,y_m\right\}$$

$$H\left(X\middle|Y\right)\neq H\left(Y\middle|X\right)$$

$$H\left(X\middle|Y=y\right)=-\sum_{i=1}^{n}p_{X|Y}\left(x_i\middle|y\right)\log p_{X|Y}\left(x_i\middle|y\right)$$

$$H\left(X\middle|Y\right)=-\sum_{j=1}^{m}p_Y\left(y_j\right)\left[\sum_{i=1}^{n}p_{X|Y}\left(x_i\middle|y_j\right)\log p_{X|Y}\left(x_i\middle|y_j\right)\right]=$$

$$=-\sum_{i=1}^{n}\sum_{j=1}^{m}p_{X,Y}\left(x_i,y_j\right)\log p_{X|Y}\left(x_i\middle|y_j\right)$$

Entropy: computation

Exemple:

$(a)\quad H(X) - ?$

$(b)\quad H(Y) - ?$

$(c)\quad H(X|Y) - ?$

$(d)\quad H(Y|X) - ?$

$(e)\quad H(X,Y) - ?$

Y \ X	1	2	3	4
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4	$\frac{1}{4}$	0	0	0

Entropy: computation

Example: basic entropy computation

$Y \backslash X$	1	2	3	4	
1	$p(X=1,Y=1)$	$p(X=2,Y=1)$	$p(X=3,Y=1)$	$p(X=4,Y=1)$	$p(Y=1)$
2	$p(X=1,Y=2)$	$p(X=2,Y=2)$	$p(X=3,Y=2)$	$p(X=4,Y=2)$	$p(Y=2)$
3	$p(X=1,Y=3)$	$p(X=2,Y=3)$	$p(X=3,Y=3)$	$p(X=4,Y=3)$	$p(Y=3)$
4	$p(X=1,Y=4)$	$p(X=2,Y=4)$	$p(X=3,Y=4)$	$p(X=4,Y=4)$	$p(Y=4)$
	$p(X=1)$	$p(X=2)$	$p(X=3)$	$p(X=4)$	

Entropy: computation

Example: basic entropy computation

$Y \backslash X$	1	2	3	4	$p(Y=i)$
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$	$p(Y=1)=\frac{1}{4}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$	$p(Y=2)=\frac{1}{4}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$p(Y=3)=\frac{1}{4}$
4	$\frac{1}{4}$	0	0	0	$p(Y=4)=\frac{1}{4}$
$p(X=j)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	

$$p(Y=1)=\sum_{j=1}^{4} p(X=j,Y=1)$$

$$p(Y=2)=\sum_{j=1}^{4} p(X=j,Y=2)$$

$$p(Y=3)=\sum_{j=1}^{4} p(X=j,Y=3)$$

$$p(Y=4)=\sum_{j=1}^{4} p(X=j,Y=4)$$

Entropy: computation

Example: basic entropy computation

$$(a) \quad H(X) = -\sum_{j=1}^{4} p(X = j) \log p(X = j)$$

$$p(X = j) = \left\{ \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8} \right\}$$

$$H(X) = -\left[\frac{1}{2}\log 2^{-1} + \frac{1}{4}\log 2^{-2} + \frac{1}{8}\log 2^{-3} + \frac{1}{8}\log 2^{-3} \right] =$$

$$= \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \frac{3}{8} = \frac{7}{4}\, bit.$$

Entropy: computation

Example: basic entropy computation

$$(b) \quad H(Y) = -\sum_{i=1}^{4} p(Y = i) \log p(Y = i)$$

$$p(Y = i) = \left\{ \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4} \right\}$$

$$H(Y) = -\left[\frac{1}{4} \log 2^{-2} + \frac{1}{4} \log 2^{-2} + \frac{1}{4} \log 2^{-2} + \frac{1}{4} \log 2^{-2} \right] =$$

$$= \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = 2 \; bit.$$

Entropy: computation

$$(c) \quad H(X|Y) = \sum_{i=1}^{4} p(Y=i) H(X|Y=i) =$$

$$= p(Y=1)H(X|Y=1) + p(Y=2)H(X|Y=2) + p(Y=3)H(X|Y=3)$$
$$+ p(Y=4)H(X|Y=4)$$

$$p(Y=i) = \left\{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right\}$$

$$H(X|Y) = \frac{1}{4}H(X|Y=1) + \frac{1}{4}H(X|Y=2) + \frac{1}{4}H(X|Y=3) + \frac{1}{4}H(X|Y=4)$$

Entropy: computation

$$H\left(X\middle|Y=1\right)=-\sum_{i=1}^{4}p\left(X=i\middle|Y=1\right)\log p\left(X=i\middle|Y=1\right)$$

Recall:

$$p(x,y)=p(x|y)p(y) \quad \Rightarrow p\left(x\middle|y\right)=\frac{p\left(x,y\right)}{p\left(y\right)}$$

From table

$$p(Y=i)=\left\{\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}\right\}$$

$$p\left(X=1\middle|Y=1\right)=\frac{p\left(X=1,Y=1\right)}{p\left(Y=1\right)}=\frac{1/8}{1/4}=\frac{1}{2}$$

$$p\left(X=2\middle|Y=1\right)=\frac{p\left(X=2,Y=1\right)}{p\left(Y=1\right)}=\frac{1/16}{1/4}=\frac{1}{4}$$

$$p\left(X=3\middle|Y=1\right)=\frac{p\left(X=3,Y=1\right)}{p\left(Y=1\right)}=\frac{1/32}{1/4}=\frac{1}{8}$$

$$p\left(X=4\middle|Y=1\right)=\frac{p\left(X=4,Y=1\right)}{p\left(Y=1\right)}=\frac{1/32}{1/4}=\frac{1}{8}$$

Entropy: computation

$$H\left(X\middle|Y=1\right)=-\sum_{i=1}^{4}p\left(X=i\middle|Y=1\right)\log p\left(X=i\middle|Y=1\right)=$$

$$=-\left[\frac{1}{2}\log 2^{-1}+\frac{1}{4}\log 2^{-2}+\frac{1}{8}\log 2^{-3}+\frac{1}{8}\log 2^{-3}\right]=$$

$$=\frac{1}{2}+\frac{1}{2}+\frac{3}{8}+\frac{3}{8}=\frac{7}{4}.$$

$$H\left(X\middle|Y=2\right)=\frac{7}{4}.\quad H\left(X\middle|Y=3\right)=2.\quad H\left(X\middle|Y=2\right)=0.$$

$$H\left(X\middle|Y\right)=\frac{1}{4}H\left(X\middle|Y=1\right)+\frac{1}{4}H\left(X\middle|Y=2\right)+\frac{1}{4}H\left(X\middle|Y=3\right)+\frac{1}{4}H\left(X\middle|Y=4\right)$$

$$=\frac{1}{4}\cdot\frac{7}{4}+\frac{1}{4}\cdot\frac{7}{4}+\frac{1}{4}\cdot 2+\frac{1}{4}\cdot 0=\frac{11}{8}\,bit$$

Entropy: computation

$$(d) \quad H(Y|X) = \sum_{j=1}^{4} p(X=j) H(Y|X=j) =$$

$$= p(X=1)H(Y|X=1) + p(X=2)H(Y|X=2) + p(X=3)H(Y|X=3)$$
$$+ p(X=4)H(Y|X=4)$$

$$p(X=j) = \left\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right\}$$

$$H(Y|X) = \frac{1}{2}H(Y|X=1) + \frac{1}{4}H(Y|X=2) + \frac{1}{8}H(Y|X=3) + \frac{1}{8}H(Y|X=4)$$

You can finalize it yourselves…

Entropy: computation

$$\Rightarrow H\left(Y\middle|X\right)=\frac{13}{8}\,bit.$$

$$H\left(X\middle|Y\right)\neq H\left(Y\middle|X\right)$$

Example: basic entropy computation

Entropy: computation

$$(e)\quad H(X,Y)=\sum_{i=1}^{4}\sum_{j=1}^{4}p(X=j,Y=i)\log p(X=j,Y=i)$$

$$=-\left[2\times\frac{1}{8}\log 2^{-3}+6\times\frac{1}{16}\log 2^{-4}+4\times\frac{1}{32}\log 2^{-5}+\frac{1}{4}\log 2^{-2}\right]=$$

$$=\frac{3}{4}+\frac{6}{4}+\frac{5}{8}+\frac{2}{4}=\frac{27}{8}bit.$$

$$H(X,Y)=H(X)+H(Y|X)$$

$$\frac{27}{8}bit.$$

$$\frac{7}{4}+\frac{13}{8}=\frac{27}{8}bit.$$

Summary: properties of entropy

- Property 1: non-negativity of entropy
- Property 2: chain rule for entropy
- Property 3: increase of entropy by “adding” new randomness
- Property 4: decrease of entropy by conditioning
- Property 5: uniform pmf has the largest entropy

Summary: properties of entropy

- Property 1: entropy is strictly positive

$$H(X) \geq 0$$

$$H(X,Y) \geq 0$$

$$H(X|Y) \geq 0$$

$$0 \leq p_X(x) \leq 1 \qquad \log p_X(x) \leq 0$$

$$0 \leq p_{X,Y}(x,y) \leq 1 \quad \Rightarrow \quad \log p_{X,Y}(x,y) \leq 0$$

$$0 \leq p_{X|Y}(x|y) \leq 1 \qquad \log p_{X|Y}(x|y) \leq 0$$

$$0 \leq -\sum p_X(x) \log p_X(x)$$

$$\Rightarrow \quad 0 \leq -\sum p_{X,Y}(x,y) \log p_{X,Y}(x,y)$$

$$0 \leq -\sum p_{X|Y}(x|y) \log p_{X|Y}(x|y)$$

Summary: properties of entropy

- Property 2: chain rule for entropy (case of 2 variables)

$$p(x,y) = p(x|y)p(y) = p(y|x)p(x) \Leftrightarrow p_{X,Y}(x,y) = p_{X|Y}(x|y)p_Y(y) = p_{Y|X}(y|x)p_X(x)$$

$$H(X,Y) = H(X|Y) + H(Y) = H(Y|X) + H(X)$$

Proof

$$-\log p_{X,Y}(x,y) = -\log p_{Y|X}(y|x)\,p_X(x) = -\log p_{Y|X}(y|x) - \log p_X(x).$$

$$H(X,Y) = E_{p_{X,Y}}\left[-\log p_{X,Y}(x,y)\right] = E_{p_{X,Y}}\left[-\log p_X(x)\right] + E_{p_{X,Y}}\left[-\log p_{Y|X}(y|x)\right]$$
$$= H(X) + H(Y|X).$$

Summary: properties of entropy

- Property 2: chain rule for entropy (case of N variables)

$$p(x_1,\cdots x_N) = p(x_1)\prod_{i=2}^{N} p(x_i \big| x_{i-1},\cdots,x_1)$$

$$H\left(X_1,X_2,...,X_N\right) = \sum_{i=1}^{N} H(X_i \big| X_{i-1},...,X_1)$$

Proof

$$H\left(X_1,X_2\right) = H\left(X_1\right) + H\left(X_2 \big| X_1\right) \qquad \Rightarrow H\left(X_1,X_2\right) \leq H\left(X_1\right) + H\left(X_2\right)$$

(=, if independent)

$$H\left(X_1,X_2,X_3\right) = H\left(X_1\right) + H\left(X_2,X_3 \big| X_1\right) = H\left(X_1\right) + H\left(X_2 \big| X_1\right) + H\left(X_3 \big| X_2,X_1\right)$$

$$H\left(X_1,X_2,...,X_N\right) = H\left(X_1\right) + H\left(X_2 \big| X_1\right) + ... + H\left(X_N \big| X_{N-1},...,X_1\right) =$$

$$= \sum_{i=1}^{N} H(X_i \big| X_{i-1},...,X_1)$$

Summary: properties of entropy

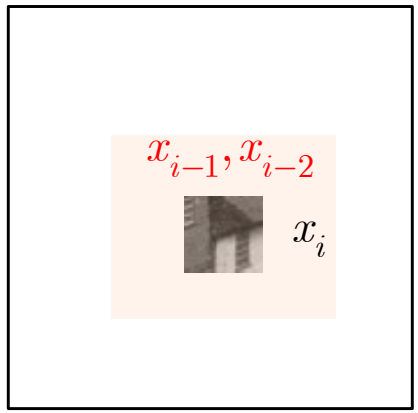
- Property 2: chain rule for entropy (case of N variables)

$$H\left(X_1, X_2, \ldots, X_N\right) = \sum_{i=1}^{N} H(X_i \left| \underbrace{X_{i-1}, \ldots, X_1}\right.)$$

$\mathbf{X} =$ 

$$\mathbf{X} \triangleq \begin{bmatrix} | \\ \mathbf{X} \\ | \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}$$

$$p_{\mathbf{x}}(\mathbf{x}) = p_{\mathbf{x}}(x_1, x_2, \cdots, x_N)$$



$$\mathbf{X} \triangleq \begin{bmatrix} | \\ \mathbf{X} \\ | \end{bmatrix} = \begin{bmatrix} \vdots \\ X_{i-2} \\ X_{i-1} \\ X_i \\ \vdots \end{bmatrix}$$

$$\prod_{i=2}^{N} p\left(x_i \left| x_{i-1}, x_{i-2} \right.\right)$$

Summary: properties of entropy

- Property 2: chain rule for entropy (case of N variables)

$$H\left(X_1, X_2, \ldots, X_N\right) \leq \sum_{i=1}^{N} H(X_i)$$

equality for independent $\{X_i\}$

Proof

For Independent $\{X_i\}$

$$H\left(X_2 \middle| X_1\right) = H\left(X_2\right)$$

$$H\left(X_3 \middle| X_2, X_1\right) = H\left(X_3\right)$$

$$H\left(X_i \middle| X_{i-1}, \ldots, X_1\right) = H\left(X_i\right)$$

$$H\left(X_1, X_2, \ldots, X_N\right) = H\left(X_1\right) + H\left(X_2 \middle| X_1\right) + \ldots + H\left(X_N \middle| X_{N-1}, \ldots, X_1\right) =$$

$$H\left(X_1, X_2, \ldots, X_N\right) = H\left(X_1\right) + H\left(X_2\right) + \ldots + H\left(X_N\right) = \sum_{i=1}^{N} H\left(X_i\right)$$

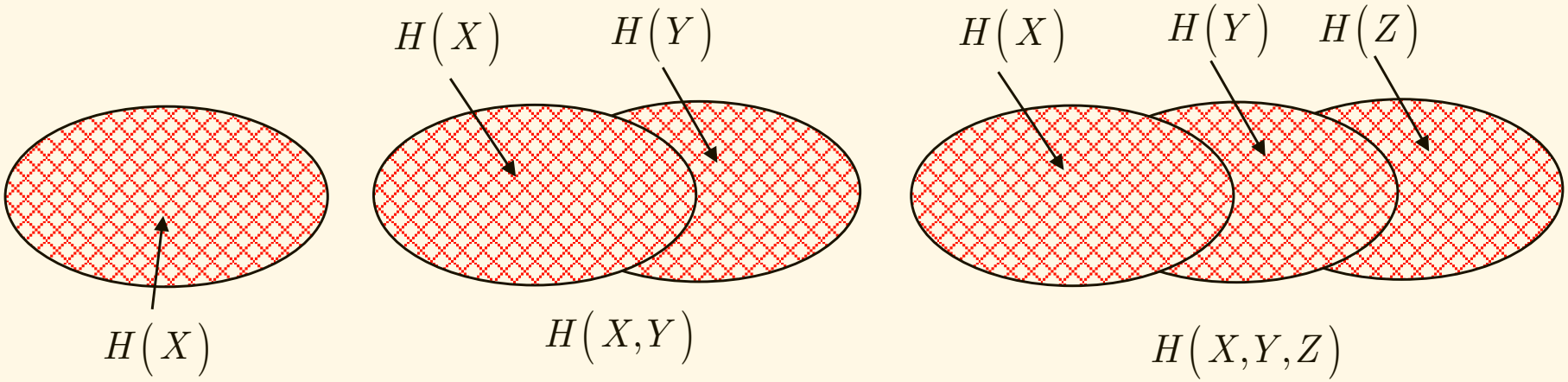
Summary: properties of entropy

- Property 3: increase of entropy by adding new randomness

$$H(X) \le H(X,Y) \le H(X,Y,Z) \le \ldots$$

$$H(X) = H(X,Y) \Leftrightarrow Y = f(X) \qquad H(X,Y) = H(X) + \underbrace{H(Y|X)}_{0}$$

$$H(X,Y) = H(X,Y,Z) \Leftrightarrow Z = f(X,Y)$$



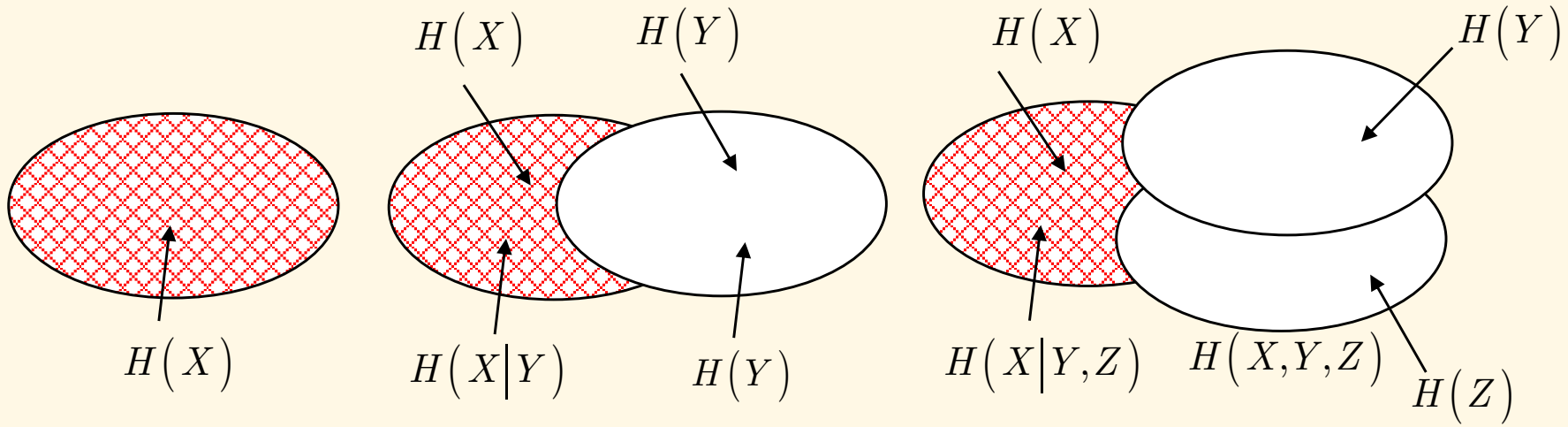
Summary: properties of entropy

- Property 4: conditioning reduces entropy

$$H(X) \geq H(X|Y) \geq H(X|Y,Z) \geq \ldots$$

$$H(X) = H(X|Y) \Leftrightarrow X \perp Y$$

$$H(X|Y) = H(X|Y,Z) \Leftrightarrow X \perp Z|Y \ldots$$



Summary: properties of entropy

- Property 5: uniform pmf has the largest entropy

$$H(X) \le \log_2 |\mathcal{X}|$$

equality, iff $\Pr[X = x_i] = \frac{1}{N} = \frac{1}{|\mathcal{X}|}$

  - Proof

Let $\mathcal{X} = \{x_1, x_2, ..., x_N\}$ and $\Pr[X = x_i] = p_i$

$$H(X) = -\sum_{x \in \mathcal{X}} \Pr[X = x] \log_2 \Pr[X = x] =$$

$$= -(p_1 \log_2 p_1 + p_2 \log_2 p_2 + ... + p_N \log_2 p_N)$$

  - Our goal: to find a distribution $\{p_i\}_{i=1}^N$ that has the maximum entropy $H(X)$

Summary: properties of entropy

- Recall: constrained optimization

$$\left(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_N\right) = \underset{x_1, x_2, \ldots, x_N: \text{ such that } g(x_1, x_2, \ldots, x_N) = k}{\arg\min} f(x_1, x_2, \ldots, x_N)$$

- We will use the method of Lagrange multipliers (unconstrained regularization)

$$\mathcal{L}(x_1, x_2, \ldots, x_N) = f(x_1, x_2, \ldots, x_N) + \lambda\Big(g(x_1, x_2, \ldots, x_N) - k\Big)$$

$$\left(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_N\right) = \underset{x_1, x_2, \ldots, x_N}{\arg\min}\, \mathcal{L}(x_1, x_2, \ldots, x_N)$$

- Solution

$$\frac{\partial \mathcal{L}(x_1, x_2, \ldots, x_N)}{\partial x_i} = \frac{\partial f(x_1, \ldots x_i, \ldots, x_N)}{\partial x_i} + \lambda \frac{\partial g(x_1, \ldots x_i, \ldots, x_N)}{\partial x_i} = 0$$

$$\frac{d\mathcal{L}(x_1, x_2, \ldots, x_N)}{d\lambda} = g(x_1, x_2, \ldots, x_N) - k = 0$$

Summary: properties of entropy

- In our case

$$f(p_1, p_2, ..., p_N) = -\sum_{i=1}^{N} p_i \log_2 p_i \qquad g(x_1, x_2, ..., x_N) = \sum_{i=1}^{N} p_i = 1$$

- The Lagrange cost

$$\mathcal{L}(x_1, x_2, ..., x_N) = f(x_1, x_2, ..., x_N) + \lambda\big(g(x_1, x_2, ..., x_N) - k\big)$$
$$\mathcal{L}(p_1, p_2, ..., p_N) = -(p_1 \log_2 p_1 + p_2 \log_2 p_2 + ... + p_N \log_2 p_N) + \lambda(p_1 + p_2 + ... + p_N - 1)$$

- Solution

$$\begin{cases} \dfrac{\partial \mathcal{L}(p_1, p_2, ..., p_N)}{\partial p_i} = -\log_2 p_i - 1 + \lambda = 0; \\ \dfrac{\partial \mathcal{L}(p_1, p_2, ..., p_N)}{\partial p_j} = -\log_2 p_j - 1 + \lambda = 0; \\ \dfrac{\partial \mathcal{L}(p_1, p_2, ..., p_N)}{\partial \lambda} = p_1 + p_2 + ... + p_N - 1 = 0; \end{cases} \Rightarrow \quad -\log_2 p_i - 1 + \lambda = -\log_2 p_j - 1 + \lambda;$$

$$\log_2 p_i = \log_2 p_j \quad \Rightarrow \quad p_i = p_j = p = \frac{1}{N}$$

$$p_i = \frac{1}{|\mathcal{X}|}; \; H_{\max}(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x) = \log_2 |\mathcal{X}|$$

Relative entropy

Definition (relative entropy): *Relative entropy* or *Kullback-Leibler divergence (KLD)* between pmfs $p(x)$ and $q(x)$:

$$D_{\mathrm{KL}}(p \parallel q) = D\left(p\middle\|q\right) = \sum_{x\in\mathcal{X}} p(x)\log_2\frac{p(x)}{q(x)} = E_{p(x)}\left[\log_2\frac{p(x)}{q(x)}\right]$$

with the conventions: $0\log_2\frac{0}{q(x)} = 0, \forall q(x)$

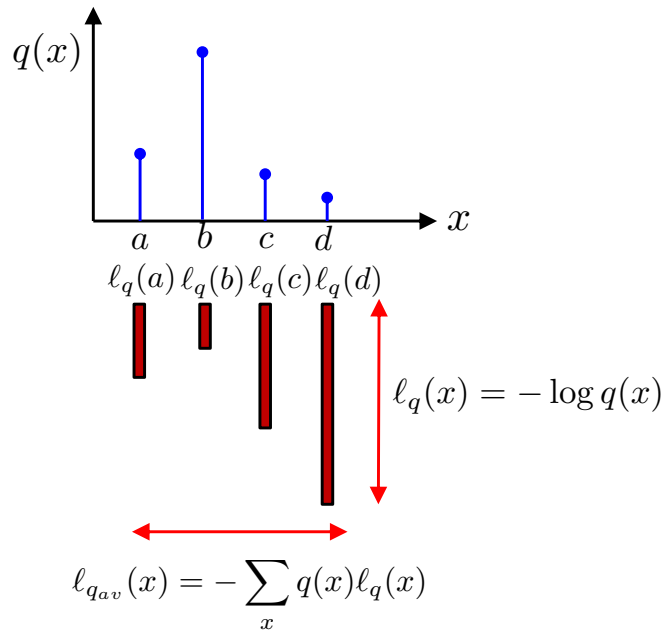
$$p(x)\log_2\frac{p(x)}{0} = +\infty, \forall p(x) > 0.$$

$$D\left(p\middle\|q\right) \neq D\left(q\middle\|p\right)$$

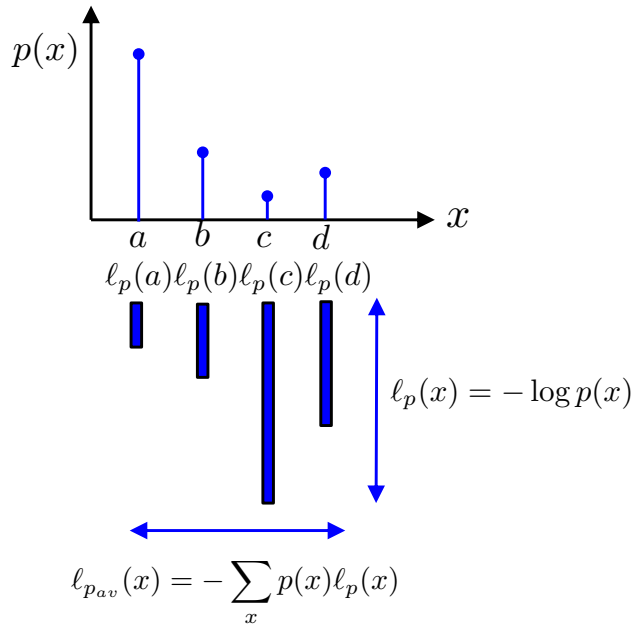
$$D\left(p\middle\|q\right) \geq 0$$

Relative entropy

Assumed data distribution $q(x)$



Data distribution $p(x)$

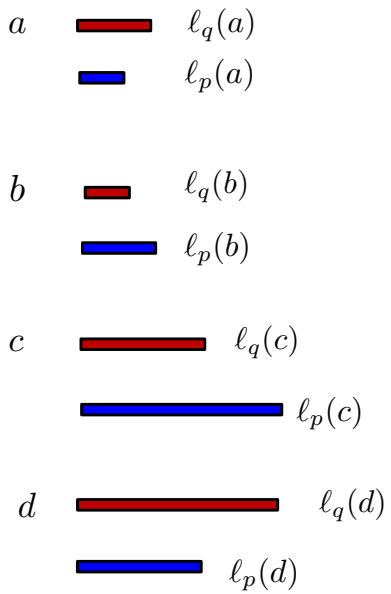


Practical problem (the true distrubution $p(x)$ is unknown)

Design code for $q(x)$ ⟶ Use the designed code for $q(x)$ to code symbols from $p(x)$

Relative entropy

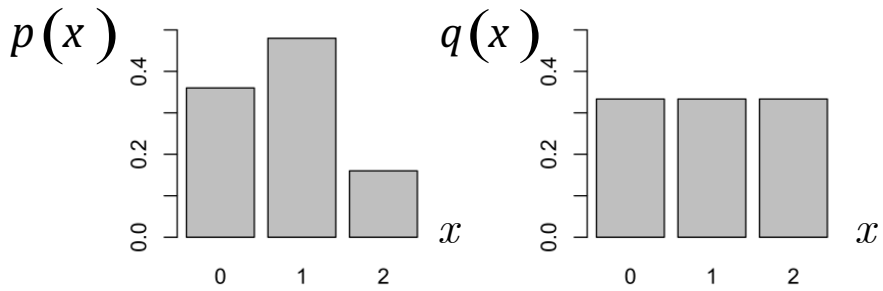
Under this assumption each symbol will be coded “suboptimally”



Question: what is the “overall loss” in this mismatched usage of codes for different sources?

Relative entropy: interpretation

- Given two distributions



- KLD "measures" in:
  - Machine learning: the level of "similarity" between these two distributions
  - Coding theory: "inefficiency" of coding designed for the pmf $q(x)$ while actually data originates from pmf $p(x)$

Real data $X \sim p(x)$

Code design $H(p(x)) + D\left(p(x)\middle\|q(x)\right)$

extra bits

$D\left(p(x)\middle\|q(x)\right) = 0$ iff $p(x) = q(x)$

Relative entropy: interpretation

- Allow fractional code lengths $\ell_q(x) = -\log q(x)$
- Then expected length for coding $X \sim p(x)$ using $\ell_q(x)$ is

$$
\begin{aligned}
L &= \mathbb{E}_{X\sim p(x)}[\ell_q(X)] \\
&= -\sum_x p(x)\log q(x) \\
&= \sum_x p(x)\log\left[\frac{p(x)}{q(x)}\frac{1}{p(x)}\right] \\
&= \sum_x p(x)\log\frac{p(x)}{q(x)} + \sum p(x)\log\frac{1}{p(x)} \\
&= D(p(x)\|q(x)) + H(p(x))
\end{aligned}
$$

→ See later as cross-etropy

Relative entropy: asymmetry

- Example:

$$x \in \mathcal{X} = \{0,1\};\ p(0) = a;\ p(1) = 1-a;\ q(0) = b;\ q(1) = 1-b.$$

$$D\left(p\middle\|q\right) = \sum_{x\in\mathcal{X}} p\left(x\right)\log_2\frac{p\left(x\right)}{q\left(x\right)} = a\log_2\frac{a}{b} + (1-a)\log_2\frac{1-a}{1-b};$$

$$D\left(q\middle\|p\right) = \sum_{x\in\mathcal{X}} q\left(x\right)\log_2\frac{q\left(x\right)}{p\left(x\right)} = b\log_2\frac{b}{a} + (1-b)\log_2\frac{1-b}{1-a}.$$

$$a = \frac{1}{4}; b = \frac{1}{8},\ D\left(p\middle\|q\right) = \frac{1}{4}\log_2\frac{8}{4} + \left(1-\frac{1}{4}\right)\log_2\left(\frac{1-\frac{1}{4}}{1-\frac{1}{8}}\right) = 0.0832 \text{ bit};$$

$$D\left(p\middle\|q\right) \neq D\left(q\middle\|p\right)$$

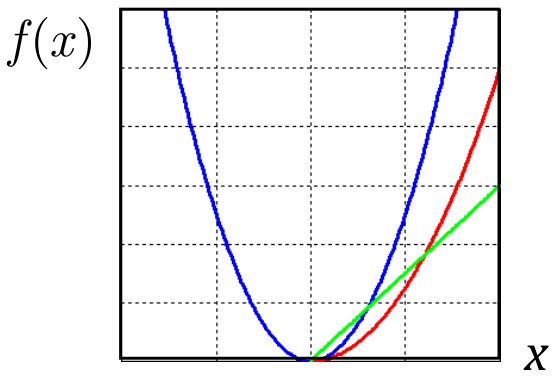
$$D\left(q\middle\|p\right) = \frac{1}{8}\log_2\frac{4}{8} + \left(1-\frac{1}{8}\right)\log_2\left(\frac{1-\frac{1}{8}}{1-\frac{1}{4}}\right) = 0.0696 \text{ bit}.$$

Relative entropy: recall of convexity

Definition (**convex function**): a function is convex (strictly convex) on the interval (a,b), if for every $x_1, x_2 \in (a,b)$ and $0 \leq \lambda \leq 1$

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

- If a function $f(x)$ has its second derivative to be non-negative (strictly positive), this function is convex (strictly convex).

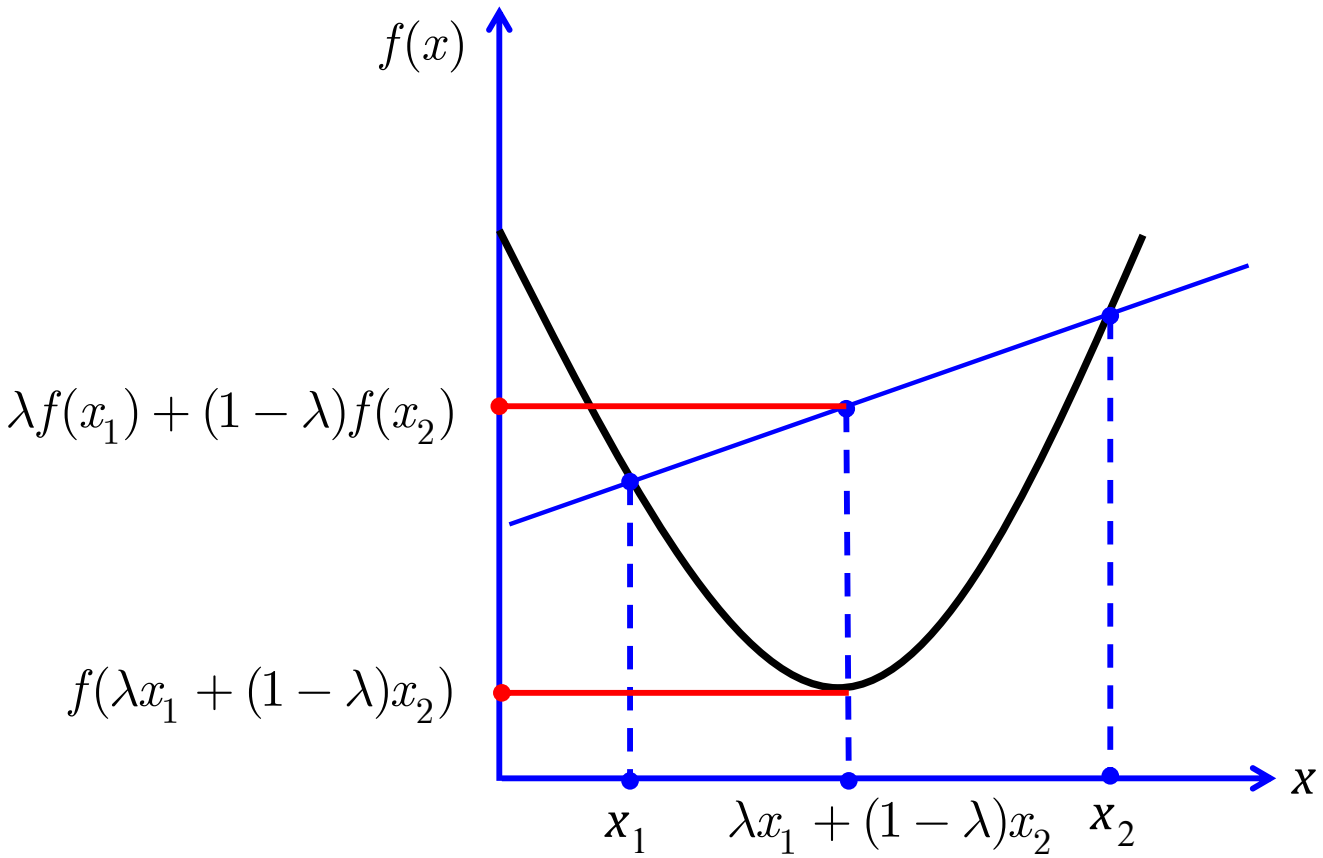


$x^2, |x|$

$e^x, x \log x$ for $x \geq 0$

Relative entropy: recall of convexity

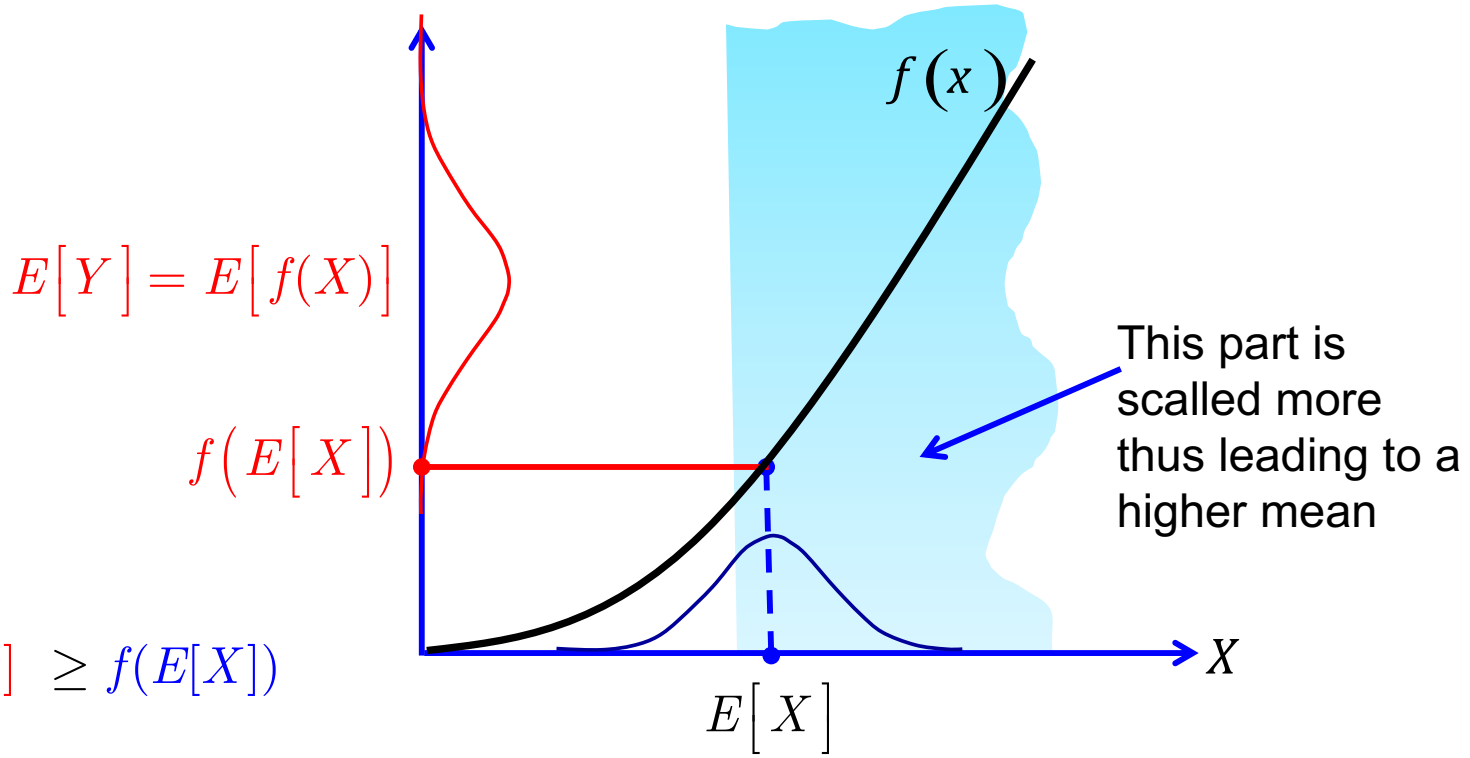
$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$



Relative entropy: Jensen's inequality

Definition (**Jensen's inequality**): for a convex function $f(x)$ and a random variable X

$$E[f(X)] \geq f(E[X])$$



$E[Y]=E[f(X)] \geq f(E[X])$

Relative entropy: log-sum inequality

Definition (**Log-sum inequality):** For nonnegative numbers $a_1, \mathrm{L}\ , a_N$ and $b_1, \cdots, b_N$

$$\sum_{i=1}^{N} a_i \log\left(\frac{a_i}{b_i}\right) \geq \left(\sum_{i=1}^{N} a_i\right) \log \frac{\sum_{i=1}^{N} a_i}{\sum_{i=1}^{N} b_i}$$

with equality if and only if $\left(\frac{a_i}{b_i}\right) = const$

Proof:

- Assume that $a_i > 0$ and $b_i > 0$
- The function $f(t) = t \log t$ is strictly convex, since $f''(t) = \frac{1}{t} \log e > 0, t > 0.$

Relative entropy: log-sum inequality

- According to Jensen's inequality, we have:

$$\sum_{i=1}^{N} \alpha_i f(t_i) \geq f\left(\sum_{i=1}^{N} \alpha_i t_i\right)$$

for $\alpha_i \geq 0, \sum_{i=1}^{N} \alpha_i = 1$

- Assume $\alpha_i = \dfrac{b_i}{\sum_{j=1}^{N} b_j}$ and $t_i = \dfrac{a_i}{b_i}$

- $$\sum_{i=1}^{N} \alpha_i f(t_i) = \sum_{i=1}^{N} \frac{b_i}{\sum_{j=1}^{N} b_j} \frac{a_i}{b_i} \log\left(\frac{a_i}{b_i}\right) = \sum_{i=1}^{N} \frac{a_i}{\sum_{j=1}^{N} b_j} \log\left(\frac{a_i}{b_i}\right)$$

Relative entropy: log-sum inequality

Proof: log-sum inequality

$$\geq f\left(\sum_{i=1}^{N}\alpha_i t_i\right) = \sum_{i=1}^{N}\alpha_i t_i \log\left(\sum_{i=1}^{N}\alpha_i t_i\right) = \sum_{i=1}^{N}\frac{b_i}{\sum_{j=1}^{N} b_j}\frac{a_i}{b_i}\log\left(\sum_{i=1}^{N}\frac{b_i}{\sum_{j=1}^{N} b_j}\frac{a_i}{b_i}\right)$$

$$= \sum_{i=1}^{N}\frac{a_i}{\sum_{j=1}^{N} b_j}\log\left(\sum_{i=1}^{N}\frac{a_i}{\sum_{j=1}^{N} b_j}\right)$$

$$\Rightarrow \sum_{i=1}^{N}\frac{a_i}{\sum_{j=1}^{N} b_j}\log\left(\frac{a_i}{b_i}\right) \geq \sum_{i=1}^{N}\frac{a_i}{\sum_{j=1}^{N} b_j}\log\left(\sum_{i=1}^{N}\frac{a_i}{\sum_{j=1}^{N} b_j}\right) \Rightarrow \sum_{i=1}^{N} a_i \log\left(\frac{a_i}{b_i}\right) \geq \sum_{i=1}^{N} a_i \log\left(\frac{\sum_{i=1}^{N} a_i}{\sum_{i=1}^{N} b_i}\right)$$

Summary: properties of relative entropy

- Property 1: non-negativity of relative entropy
- Property 2: conditional relative entropy and chain rule
- Property 3: relative entropy wrt uniform pdf

Summary: properties of relative entropy

- Property 1: non-negativity of relative entropy

$$D_{\mathrm{KL}}\left(p\middle\|q\right) \geq 0$$

$$\text{"}=\text{"},\ \text{iff}\ p\left(x\right)=q\left(x\right),\ \forall x\in\mathcal{X}$$

- Proof: $S=\{x : p(x)>0\}$

$$-D_{\mathrm{KL}}(p\|q) = -\sum_{x\in S} p(x)\log\left(\frac{p(x)}{q(x)}\right) = \sum_{x\in S} p(x)\log\left(\frac{q(x)}{p(x)}\right) \leq$$

Log-sum INQ

$$\sum_i a_i \log\left(\frac{a_i}{b_i}\right) \geq \left(\sum_i a_i\right)\log\frac{\sum_i a_i}{\sum_i b_i}$$

$$\leq \left(\sum_{x\in S} p(x)\right)\log\frac{\sum_{x\in S}\left(q(x)\right)}{\sum_{x\in S}\left(p(x)\right)} = 1\cdot\log\frac{1}{1} = 0$$

$$\boxed{D_{\mathrm{KL}}(p\|q)\geq 0}$$

$$D_{\mathrm{KL}}(p\|q)=0$$
$$p(x)=q(x)$$

$$-D(p(x)\|q(x)) = \sum p(x)\log\left(\frac{q(x)}{p(x)}\right) \leq \log\left(\sum p(x)\frac{q(x)}{p(x)}\right) = \log\left(\sum q(x)\right) = 0$$

Summary: properties of relative entropy

- Property 2: conditional relative entropy and chain rule
  - Conditional relative entropy

$$D_{\mathrm{KL}}\left(p\left(y\middle|x\right)\middle\|q\left(y\middle|x\right)\right)=\sum_{x\in\mathcal{X}}p\left(x\right)\sum_{y\in\mathcal{Y}}p\left(y\middle|x\right)\log_2\frac{p\left(y\middle|x\right)}{q\left(y\middle|x\right)}=E_{p\left(x,y\right)}\left[\log_2\frac{p\left(y\middle|x\right)}{q\left(y\middle|x\right)}\right]$$

  - Chain rule for relative entropy

$$D_{\mathrm{KL}}\left(p\left(x,y\right)\middle\|q\left(x,y\right)\right)=D_{\mathrm{KL}}\left(p\left(x\right)\middle\|q\left(x\right)\right)+D_{\mathrm{KL}}\left(p\left(y\middle|x\right)\middle\|q\left(y\middle|x\right)\right)$$

Summary: properties of relative entropy

- Proof:

$$D_{\mathrm{KL}}\left(p(x,y)\middle\|q(x,y)\right)=\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}p(x,y)\log_2\frac{p(x,y)}{q(x,y)}$$

$$=\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}p(x,y)\log_2\frac{p(x)}{q(x)}\frac{p(y|x)}{q(y|x)} \qquad \left| \; p(x,y)=p(x)p(y|x)\right.$$

$$=\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}p(x,y)\log_2\frac{p(x)}{q(x)}+\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}p(x,y)\log_2\frac{p(y|x)}{q(y|x)} \qquad \left| \; \log ab=\log a+\log b\right.$$

$$=D_{\mathrm{KL}}\left(p(x)\middle\|q(x)\right)+D_{\mathrm{KL}}\left(p(y|x)\middle\|q(y|x)\right)\ \square$$

Summary: properties of relative entropy

- Property 3: relative entropy wrt uniform pmf
  - Assume a random variable with the **uniform distribution** $\left(q\left(x\right)=1/\left|\mathcal{X}\right|\right)$
  - Then, KLD between any $p(x)$ and uniform $q(x)$ is:

$$D_{\mathrm{KL}}\left(p\middle\|q\right)=\log_2\left|\mathcal{X}\right|-H\left(X\right)$$

$$D_{\mathrm{KL}}\left(p\middle\|q\right)=\sum_{x\in\mathcal{X}}p\left(x\right)\log_2\frac{p\left(x\right)}{q\left(x\right)}=\underbrace{\sum_{x\in\mathcal{X}}p\left(x\right)\log_2 p\left(x\right)}_{-H\left(X\right)}-\underbrace{\sum_{x\in\mathcal{X}}p\left(x\right)}_{1}\log_2\underbrace{q\left(x\right)}_{1/\left|\mathrm{X}\right|}$$

$$\Rightarrow H\left(X\right)=\log_2\left|\mathcal{X}\right|-D_{\mathrm{KL}}\left(p\middle\|q\right)$$

- Thus, the distribution $p(x)$ that has the maximum entropy $H(X)=\log_2\left|\mathcal{X}\right|$ is uniform, i.e., since $D_{\mathrm{KL}}\left(p\middle\|q\right)\geq 0$

Cross-entropy

Definition (**cross-entropy**): *Cross-entropy* between pmfs $p(x)$ and $q(x)$:

$$H(p,q) = -\sum_{x\in\mathcal{X}} p(x)\log q(x) = -\mathrm{E}_{p(x)}[\log q(x)]$$

- A link to KLD:

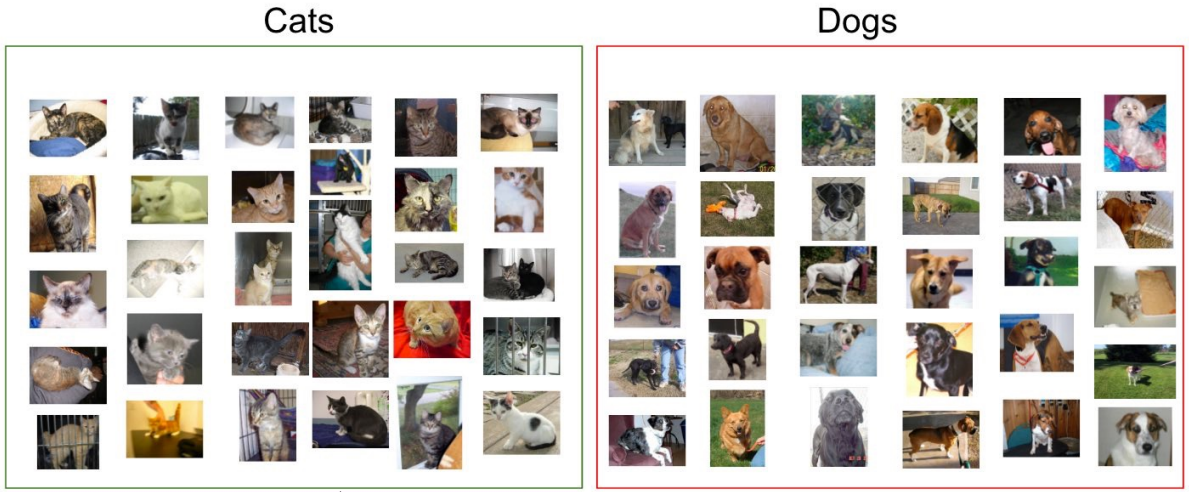
$$H(p,q) = H(p) + D_{\mathrm{KL}}(p \parallel q)$$

  - Proof:

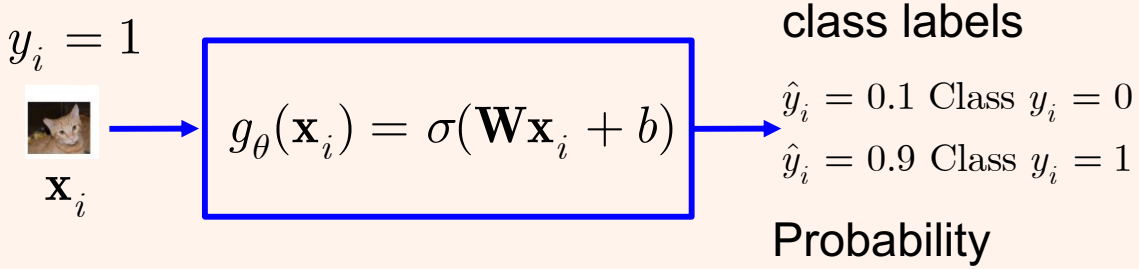
$$\begin{aligned} D_{\mathrm{KL}}(p \parallel q) &= E_{p(x)}\left[\log\frac{p(x)}{q(x)}\right] = E_{p(x)}\left[\log p(x)\right] - E_{p(x)}\left[\log q(x)\right] \\ &= -H(p) + H(p;q) \end{aligned}$$

- Remark: since $H(p)$ is constant, optimizing $H(p,q)$ wrt $q(x)$ is equivalent to $D_{\mathrm{KL}}(p \parallel q)$

Cross-entropy: in binary classification



- Total N training samples
- All samples are labeled $\{\mathbf{x}_i, y_i\}_{i=1}^{N}$
- We have two labels $y_i \in \{0,1\}$



- Goal:

Train the classifier $g_\theta(\mathbf{x}_i)$ producing the most accurate class prediction

Cross-entropy: in binary classification

- Notations:
  - True label $y_i \in \{0,1\}$
  - Precited label $\hat{y}_i = p_\theta\left(y_i = 1 \mid \mathbf{x}_i\right) = g_\theta\left(\mathbf{x}_i\right)$
    $1 - \hat{y}_i = p_\theta\left(y_i = 0 \mid \mathbf{x}_i\right) = 1 - g_\theta\left(\mathbf{x}_i\right)$

- Log-likelihood for N independent training samples:

$$\ell(\theta) = \log L(\theta) = \log \prod_{i=1}^{N} \underbrace{\left[g_\theta\left(\mathbf{x}_i\right)\right]^{\left(y_i\right)}}_{\hat{y}_i} \underbrace{\left[1 - g_\theta\left(\mathbf{x}_i\right)\right]^{\left(1-y_i\right)}}_{1-\hat{y}_i}$$

$$= \sum_{i=1}^{N} y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)$$

- Optimal estimation of parameter according to Maximum Likelihood (ML)

$$\hat{\theta} = \arg\max_\theta \ell(\theta) = \arg\min_\theta -\ell(\theta) = \arg\min_\theta \sum_{i=1}^{N} \underbrace{-\left(y_i \log \hat{y}_i + (1 - y_i) \log\left(1 - \hat{y}_i\right)\right)}_{H(y_i;\hat{y}_i)}$$

Cross-entropy: in binary classification

- Summary:
  - True label pmf $p \in \{y, 1-y\}$
  - Predicted label pmf $q \in \{\hat{y}, 1-\hat{y}\}$

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N} H\left(y_i, \hat{y}_i\right) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1-y_i)\log\left(1-\hat{y}_i\right)\right]$$

- Link to KLD (entropy is constant and it is equivalent)

$$H(p,q) = \underbrace{H(p)}_{const} + D_{\mathrm{KL}}(p \parallel q)$$

Cross-entropy: in binary classification

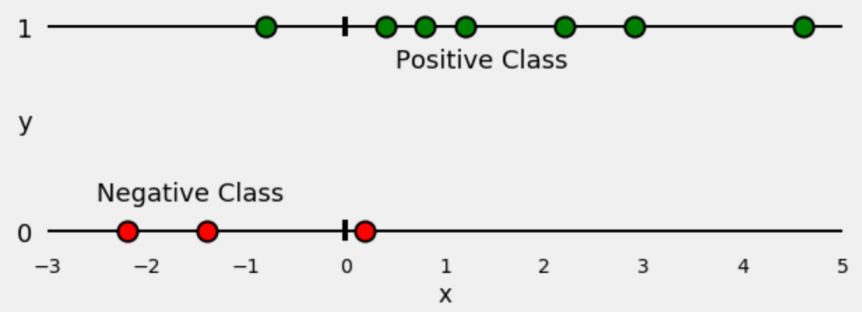


Figure 2: splitting the data!

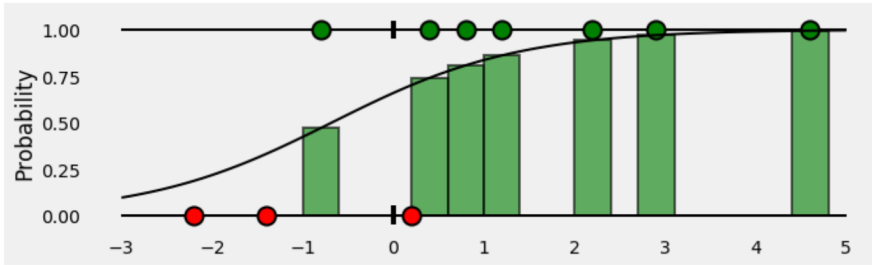


Figure 4: probabilities of classifying points in the POSITIVE class correctly

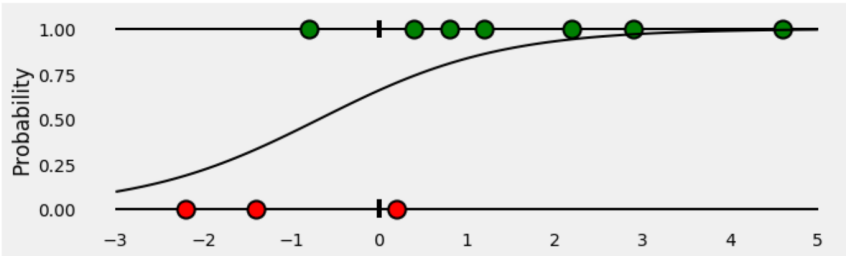


Figure 3: fitting a Logistic Regression

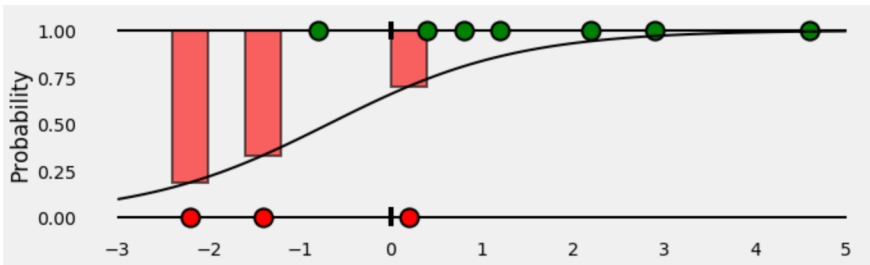


Figure 5: probabilities of classifying points in the NEGATIVE class correctly

https://towardsdatascience.com/understanding-binary-cross-entropy-log-loss-a-visual-explanation-a3ac6025181a

Mutual information

Definition (mutual information): *Mutual information* between two r.v. X and Y is defined by:

$$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}(x,y)\log\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} = E_{p_{X,Y}(x,y)}\left[\log\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}\right]$$

$$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}(x,y)\log\frac{p_{X|Y}(x|y)\,p_Y(y)}{p_X(x)\,p_Y(y)}$$

$$I(X;Y) = I(Y;X)$$

$$I(X;Y) \geq 0$$

$$\Rightarrow I(X;Y) = 0 \Leftrightarrow X \perp Y \quad \text{(if independent)}$$

Mutual information

Mutual information via KL-divergence

$$I\left(X;Y\right) = D_{\mathrm{KL}}\left(p_{X,Y}\left(x,y\right)\middle\|p_X\left(x\right)p_Y\left(y\right)\right) = E_{p_{X,Y}\left(x,y\right)}\left[\log\frac{p_{X,Y}\left(x,y\right)}{p_X\left(x\right)p_Y\left(y\right)}\right]$$

$$I(X;Y) = \mathbb{E}_{p(y)}\left[D_{\mathrm{KL}}(p(x|y)\|p(x))\right] = \mathbb{E}_{p(x)}\left[D_{\mathrm{KL}}(p(y|x)\|p(y))\right]$$

- Mutual information measures how different the joint pmf is from the product of marginal pmfs, i.e., a real pmf, vs an assumption of independent variables

Pay attention: on the equivalence of notations $p_{X,Y}\left(x,y\right) = p\left(x,y\right)$

Mutual information

Mutual information via KL-divergence

$$I(X;Y) = D_{\mathrm{KL}}\left(p(x,y) \| p(x)p(y)\right) = \mathbb{E}_{p(x,y)}\left[\log \frac{p(x,y)}{p(x)p(y)}\right]$$

$$= \mathbb{E}_{p(x)p(y|x)}\left[\log \frac{p(y \mid x)p(x)}{p(x)p(y)}\right] = \mathbb{E}_{p(x)}\left[\mathbb{E}_{p(y|x)}\left[\log \frac{p(y \mid x)}{p(y)}\right]\right]$$

$$= \mathbb{E}_{p(x)}\left[D_{\mathrm{KL}}(p(y \mid x) \| p(y))\right]$$

Summary: properties of mutual information

- Property 1: non-negativity of mutual information
- Property 2: mutual information decomposition via entropy
- Property 3: conditional mutual information
- Property 4: chain rule for mutual information
- Property 5: non-creativity of information by processing /data processing inequality/

Mutual information: properties

- Property 1: non-negativity of mutual information (from non-negativity of KLD)

$$I(X;Y) \geq 0, \quad I(X;Y,Z) \geq 0$$

$$\Rightarrow I(X;Y) = 0 \Leftrightarrow X \perp Y$$

If X and Y are independent

$$I(X;Y) = H(X) - H(X|Y) \qquad H(X) \geq 0, H(X|Y) \geq 0$$

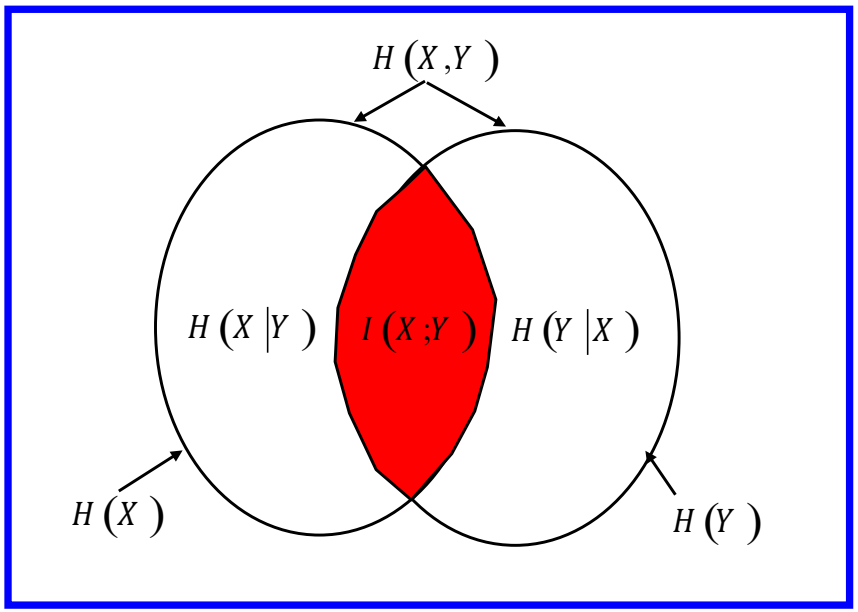
$$H(X) \geq H(X|Y) \Rightarrow I(X;Y) \geq 0$$

- Property 2: mutual information decomposition via entropy

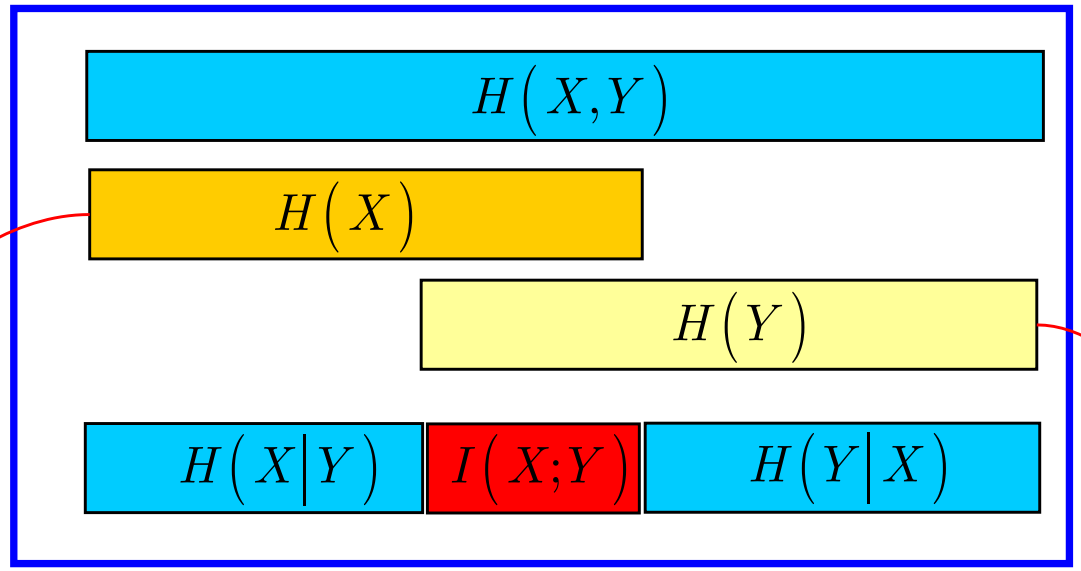
$$I(X;Y) = H(X) - H(X|Y)$$
$$I(X;Y) = H(Y) - H(Y|X)$$
$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

$$I(X;X) = H(X)$$

$$I(X;X) = H(X) - \underbrace{H(X|X)}_{0}$$



Mutual information: properties



$$I(X;Y) = H(X) - H(X|Y) \qquad I(X;Y) = H(Y) - H(Y|X)$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

Démo:

$$I(X;Y) = E_{p_{X,Y}}\left[\log \frac{p_{X|Y}(x|y)}{p_X(x)}\right] = \underbrace{E_{p_{X,Y}}\left[\log p_{X|Y}(x|y)\right]}_{-H(X|Y)} - \underbrace{E_{p_{X,Y}}\left[\log p_X(x)\right]}_{H(X)}$$

Summary: properties of mutual information

- Property 3: conditional mutual information

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$$

$$I(X;Y|Z) = E_{p(x,y,z)}\left[\log \frac{p_{X,Y|Z}(x,y|z)}{p_{X|Z}(x|z)\,p_{Y|Z}(y|z)}\right] = E_{p(x,y,z)}\left[\log \frac{p_{X|Y,Z}(x|y,z)}{p_{X|Z}(x|z)}\right]$$

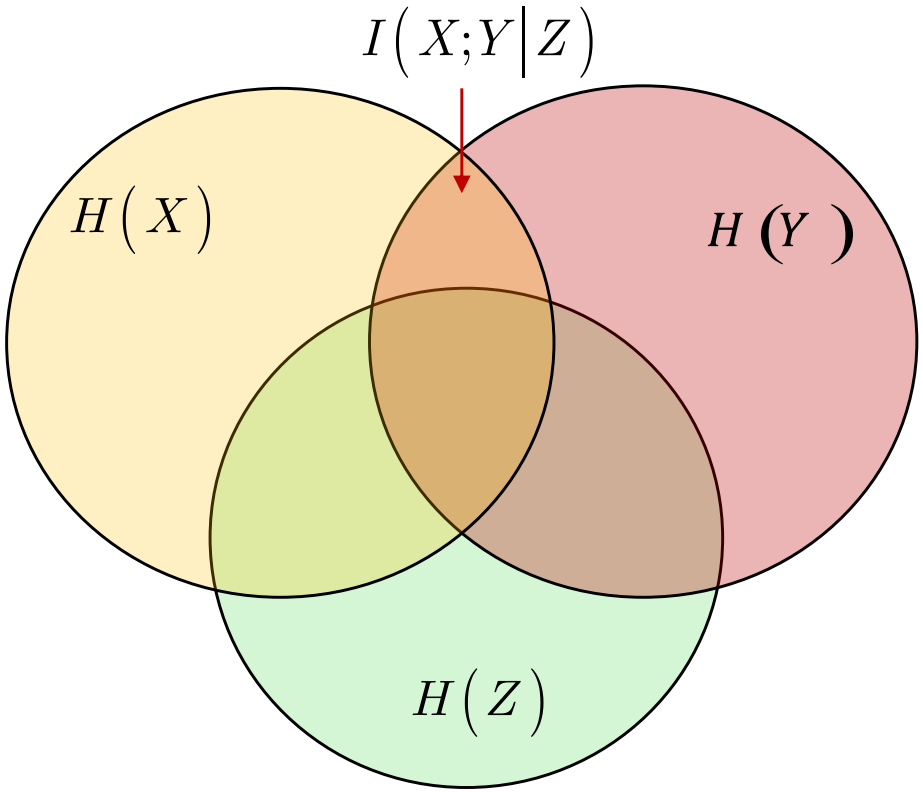
$$p_{X,Y|Z}(x,y|z) = p_{Y|Z}(y|z)\,p_{X|Y,Z}(x|y,z)$$

$$= \underbrace{E_{p(x,y,z)}\left[\log p_{X|Y,Z}(x|y,z)\right]}_{-H(X|Y,Z)} - \underbrace{E_{p(x,y,z)}\left[\log p_{X|Z}(x|z)\right]}_{H(X|Z)}$$

Summary: properties of mutual information

- Property 3: conditional mutual information

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$$



Summary: properties of mutual information

- Property 4: chain rule for mutual information

$$I\left(X_1, X_2, \ldots, X_N; Y\right) = \sum_{i=1}^{N} I(X_i; Y \mid X_{i-1}, \ldots, X_1)$$

$$I\left(X_1, X_2, \ldots, X_N; Y\right) = H\left(X_1, X_2, \ldots, X_N\right) - H\left(X_1, X_2, \ldots, X_N \mid Y\right)$$

$$\left\{ I\left(Z; Y\right) = H\left(Z\right) - H\left(Z \mid Y\right) \right\}$$

Chain rule

$$\left\{ H(X_1, X_2, \ldots, X_N) = \sum_{i=1}^{N} H(X_i \mid X_{i-1}, \ldots, X_1) \right\}$$

$$= \sum_{i=1}^{N} H(X_i \mid X_{i-1}, \ldots, X_1) - \sum_{i=1}^{N} H(X_i \mid X_{i-1}, \ldots, X_1, Y) =$$

$$= \sum_{i=1}^{N} I(X_i; Y \mid X_{i-1}, \ldots, X_1)$$

Summary: properties of mutual information

- Property 4: chain rule for mutual information

$$I\left(X_1, X_2, \ldots, X_N; Y\right) = \sum_{i=1}^{N} I(X_i; Y \,|\, X_{i-1}, \ldots, X_1)$$

Image pixels

$$\mathbf{X} = \begin{bmatrix} | \\ \mathbf{x} \\ | \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix} \in \mathcal{X}^N$$

$N = n \times m$

Label

$$Y \in \mathcal{Y}$$

Mutual information: properties

- Example: more complex decomposition

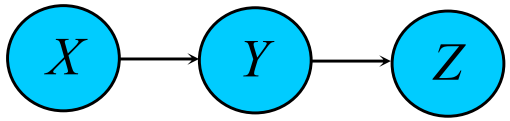
$$I\left(X_1, X_2; Y_1, Y_2\right) = I\left(X_1; Y_1, Y_2\right) + I\left(X_2; Y_1, Y_2 \middle| X_1\right)$$

$$I\left(X_1; Y_1, Y_2\right) = I\left(X_1; Y_1\right) + I\left(X_1; Y_2 \middle| Y_1\right) = I\left(X_1; Y_2\right) + I\left(X_1; Y_1 \middle| Y_2\right)$$

$$\begin{aligned} I\left(X_2; Y_1, Y_2 \middle| X_1\right) &= I\left(X_2; Y_1 \middle| X_1\right) + I\left(X_2; Y_2 \middle| X_1, Y_1\right) \\ &= I\left(X_2; Y_2 \middle| X_1\right) + I\left(X_2; Y_1 \middle| X_1, Y_2\right) \end{aligned}$$

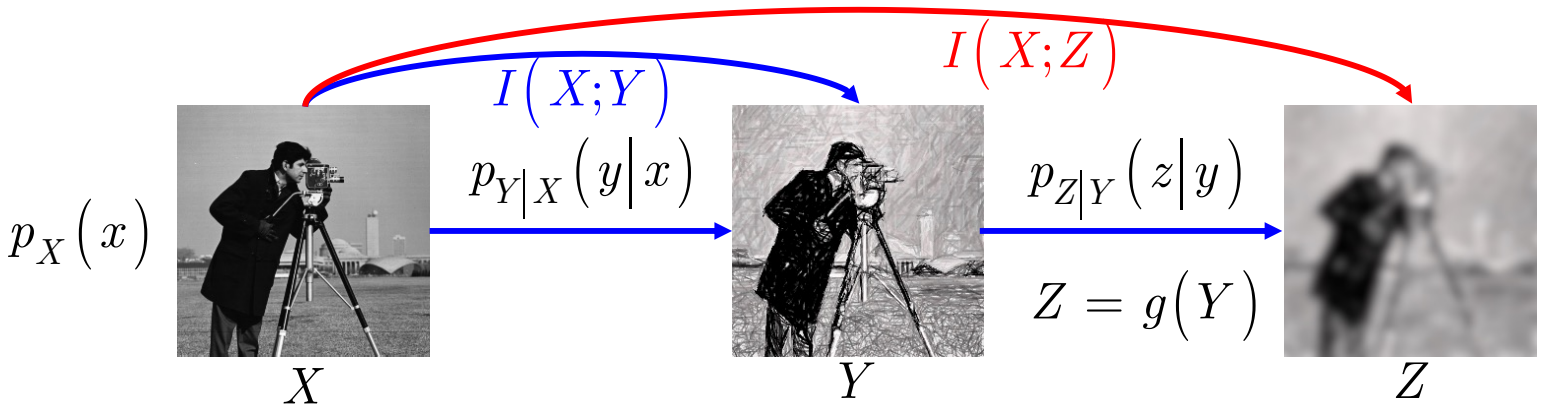
Summary: properties of mutual information

- Property 5: non-creativity of information by processing /data processing inequality/
  - Consider the Markov chain $p_{X,Y,Z}\left(x,y,z\right)=p_X\left(x\right)p_{Y|X}\left(y\middle|x\right)p_{Z|Y}\left(z\middle|y\right)$



  - Two inequalities are valid

$$I\left(X;Y\right)\geq I\left(X;Z\right)\qquad I\left(Y;Z\right)\geq I\left(X;Z\right)$$



Summary: properties of mutual information

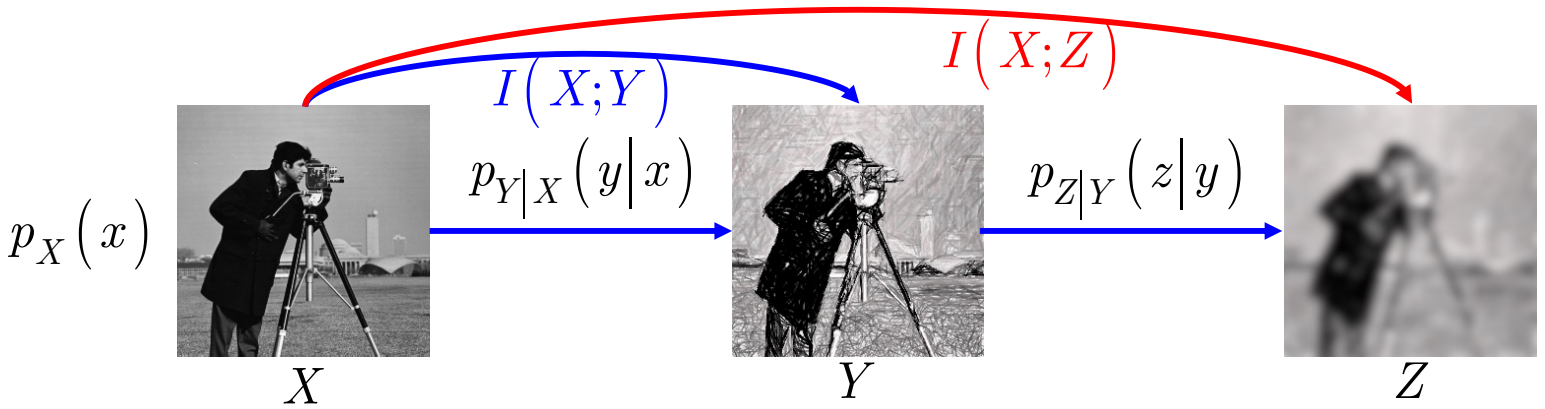
- **Data processing inequality**

  The amount of information cannot be increased by any processing $Z = g(Y)$

  This also includes deep networks!

- Two inequalities are valid

$$I(X;Y) \geq I(X;Z) \qquad I(Y;Z) \geq I(X;Z)$$



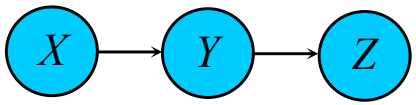
Summary: properties of mutual information

- **Proof of data processing inequality**
  - Using chain rule for mutual information

$$I\left(X;Y,Z\right) = I\left(X;Z\right) + I\left(X;Y\middle|Z\right) \quad (a)$$

$$= I\left(X;Y\right) + I\left(X;Z\middle|Y\right) \quad (b)$$

$$I\left(X;Z\middle|Y\right) = 0$$

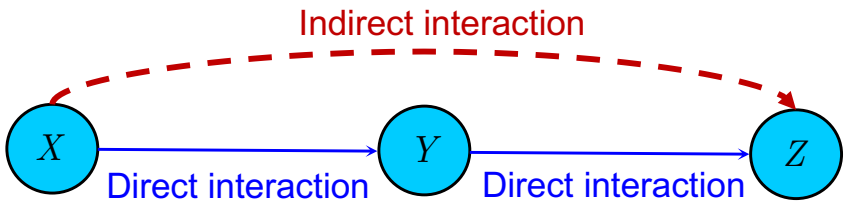


$$X \perp Z\middle|Y$$

$$I\left(X;Y\right) = I\left(X;Z\right) + I\left(X;Y\middle|Z\right) \quad \Rightarrow \quad I\left(X;Y\right) \geq I\left(X;Z\right)$$

Summary: properties of mutual information

- Consider a general case

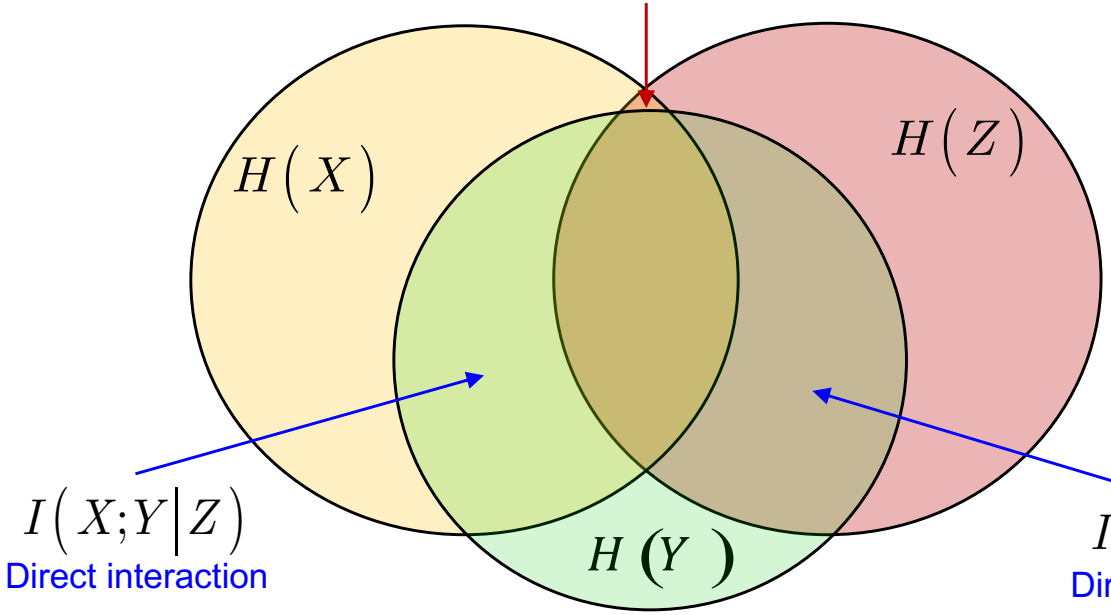


$I\left(X;Z\middle|Y\right)\to 0$ depending on the form: for Markov

Indirect interaction

$X \perp Z \middle| Y$

$I\left(X;Z\middle|Y\right)=0$



$I\left(X;Y\middle|Z\right)$
Direct interaction

$I\left(Y;Z\middle|X\right)$
Direct interaction

Scope

Part II: continuous random variables and vectors

Scope

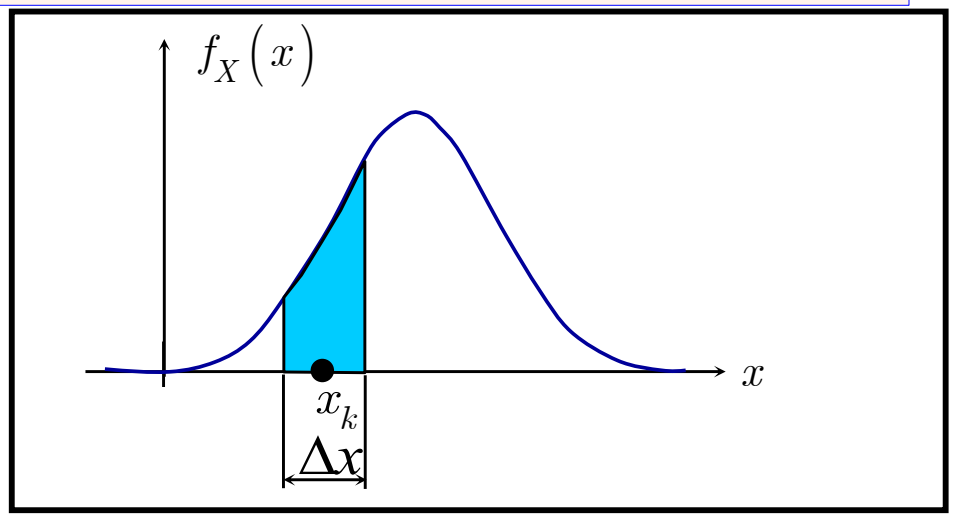
- Information theoretic measures for discrete/continuous variables
  - Entropy → Differential entropy
  - Conditional entropy
  - Joint entropy
  - Relative entropy (KL-divergence)
  - Cross entropy
  - Mutual information
  - Additional topics:
    - f-divergence
    - variational inference
    - practical computations from samples

Continuous R.Vs.: Differential entropy

Definition (**entropy**): *Differential entropy* of r.v. X with pdf $f_X(x)$

$$h(X) = -\int_{\mathcal{X}} f_X(x)\log_2 f_X(x)\,dx = E_{f_X}\left[-\log_2 f_X(x)\right]$$

$$P\left[x_k;\Delta x\right] = \int_{\Delta x} f_X(x)dx \cong f_X(x_k)\Delta x$$



Continuous R.Vs.: Differential entropy

$$H\left(X;\Delta x\right)=-\sum_{i=1}^{N}P\left[x_k,\Delta x\right]\log_2 P\left[x_k,\Delta x\right]=-\sum_{i=1}^{N}f_X\left(x_k\right)\Delta x\log_2 f_X\left(x_k\right)\Delta x$$

$$=-\sum_{i=1}^{N}f_X\left(x_k\right)\Delta x\log_2 f_X\left(x_k\right)-\underbrace{\sum_{i=1}^{N}f_X\left(x_k\right)\Delta x}_{=\,1}\underbrace{\log_2\Delta x}_{const}$$

$$\int_{-\infty}^{+\infty}f_X\left(x\right)dx=1$$

In the limit Δx tends to zero for large N. As a result, $\log_2\Delta x$ tends to infinity.

$$H\left(X;\Delta x\right)=\underbrace{-\int f_X\left(x\right)\log_2 f_X\left(x\right)dx}_{h(X)}-\log_2\Delta x.$$

Continuous R.Vs.: Differential entropy

Definition (Differential entropy of Gaussian r.v.): *differential entropy of Gaussian r.v.* $X \sim \mathcal{N}\left(0, \sigma_X^2\right)$ is:

$$h\left(X\right) = -E_{f_X}\left[\log_2 f_X\left(x\right)\right] = \frac{1}{2}\log_2\left(2\pi e\sigma_X^2\right)$$

- Proof:

$$h\left(X\right) = -\int_{\mathcal{X}} f_X\left(x\right)\ln f_X\left(x\right)dx \quad [nats] = -\int_{-\infty}^{\infty} f_X\left(x\right)\left[-\frac{x^2}{2\sigma_X^2} - \ln\sqrt{2\pi\sigma_X^2}\right]dx =$$

$$= \int_{-\infty}^{\infty} f_X\left(x\right)\frac{x^2}{2\sigma_X^2}dx + \ln\sqrt{2\pi\sigma_X^2}\int_{-\infty}^{\infty} f_X\left(x\right)dx$$

Continuous R.Vs.: Differential entropy

$$h(X) = \int_{-\infty}^{\infty} f_X(x) \frac{x^2}{2\sigma_X^2} dx + \ln\sqrt{2\pi\sigma_X^2} \underbrace{\int_{-\infty}^{\infty} f_X(x)\, dx}_{=1}$$

$$Var[X] = \int_{-\infty}^{\infty} x^2 f_X(x)\, dx = \sigma_X^2$$

$$\int_{-\infty}^{\infty} f_X(x) \frac{x^2}{2\sigma_X^2} dx = \frac{\sigma_X^2}{2\sigma_X^2} = \frac{1}{2}$$

$$h(X) = \frac{1}{2} + \ln\sqrt{2\pi\sigma_X^2} = \frac{1}{2}\ln e + \frac{1}{2}\ln 2\pi\sigma_X^2 = \frac{1}{2}\left[\ln e + \ln 2\pi\sigma_X^2\right] = \frac{1}{2}\ln\left(2\pi e\sigma_X^2\right)[nats]$$

$$h(X) = -E_{f_X}\left[\log_2 f_X(x)\right] = \frac{1}{2}\log_2\left(2\pi e\sigma_X^2\right)[\text{bits}]$$

Continuous R.Vs.: Differential entropy

- Differential entropy can be negative

$$h(X) = \frac{1}{2}\log_2\left(2\pi e\sigma_X^2\right)$$

$$\log_2(a) \leq 0, \text{ for } a \leq 1$$

$$h(X) \leq 0, \text{ for } 2\pi e\sigma_X^2 \leq 1 \quad \Rightarrow \quad \sigma_X^2 \leq \frac{1}{2\pi e}$$

- Entropy of discrete random variables $H(X) \geq 0$

Continuous R.Vs.: Differential entropy of sum

Definition (Differential entropy of sum of two independent Gaussian r.v.s): given two independent random variables $X_1 \sim \mathcal{N}\left(\overline{x}_1, \sigma^2_{X_1}\right)$ and $X_2 \sim \mathcal{N}\left(\overline{x}_2, \sigma^2_{X_2}\right)$, their sum $Y = X_1 + X_2$ has the differential entropy

$$h\left(Y\right) = E_{f_Y}\left[-\log_2 f_Y\left(y\right)\right] = \frac{1}{2}\log_2\left(2\pi e\left(\sigma^2_{X_1} + \sigma^2_{X_2}\right)\right)$$

Continuous R.Vs.: Differential entropy

Definition (Differential entropy of uniform r.v.): *differential entropy of uniform r.v.* X defined on $\left[a,b\right]$ is:

$$h\left(X\right)=\log_2(b-a)$$

- Proof:

$$h\left(X\right)=-\int_{\mathcal{X}} f_X\left(x\right)\log_2 f_X\left(x\right)dx=-\int_a^b \frac{1}{b-a}\log_2\frac{1}{b-a}dx=\log_2(b-a)$$

Continuous R.Vs.: Differential entropy

Definition (Differential entropy of exponential r.v.): *differential entropy of exponential r.v.* X is:

$$h(X) = \log_2 \frac{e}{\lambda}$$

- Proof:

$$h(X) = -\int_{\mathcal{X}} f_X(x) \log_2 f_X(x) dx = -\int_0^\infty \lambda \exp(-\lambda x) \log_2(\lambda \exp(-\lambda x)) dx =$$

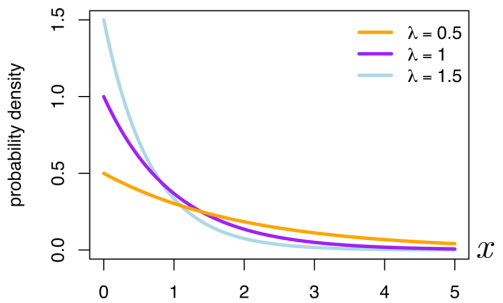
$$= -\int_0^\infty \lambda \exp(-\lambda x)(\log_2(\lambda) - \lambda x \log_2(e))) dx = I_1 + I_2;$$

$$I_1 = -\log_2(\lambda) \underbrace{\int_0^\infty \lambda \exp(-\lambda x) dx}_{=1} = \log_2 \frac{1}{\lambda};$$

$$I_2 = -\int_0^\infty -(\lambda x \log_2(e)) \lambda \exp(-\lambda x) dx = \lambda \log_2(e) \underbrace{\int_0^\infty x \lambda \exp(-\lambda x) dx}_{E[X]=1/\lambda} =$$

$$= \lambda \log_2(e) E[X] = \log_2(e);$$

$$I_1 + I_2 = \log_2 \frac{1}{\lambda} + \log_2(e) = \log_2 \frac{e}{\lambda} = 1 - \ln \lambda$$



Continuous R.Vs.: Conditional differential entropy

Definition (Conditional differential entropy): *conditional differential entropy* is defined as

$$h\left(X\middle|Y\right)=-\int_{\mathcal{X}}\int_{\mathcal{Y}}f_{X,Y}\left(x,y\right)\log_2 f_{X|Y}\left(x\middle|y\right)dxdy=h\left(X,Y\right)-h\left(Y\right)$$

$$f_{X|Y}\left(x\middle|y\right)=f_{X,Y}\left(x,y\right)/\,f_Y\left(y\right)$$

$$h\left(X\middle|Y\right)\neq h\left(Y\middle|X\right)$$

Continuous R.Vs.: Joint differential entropy

Definition (Joint differential entropy): *joint differential entropy* is defined as

$$h(X,Y) = -\int_{\mathcal{X}}\int_{\mathcal{Y}} f_{X,Y}(x,y)\log_2 f_{X,Y}(x,y)dxdy = \\ = h(X) + h(Y|X) = h(Y) + h(X|Y)$$

$$h(X,Y) = h(Y,X)$$

Continuous R.Vs.: max entropy for fixed variance

- Gaussian has maximum differential entropy among all distributions with the same variance

- Let $q(x)$ be any density satisfying $\int x^2 q(x)\,dx = \sigma^2$. Let $p(x) = \mathcal{N}\left(0, \sigma_X^2\right)$. Then

$$h(q(x)) \leq h(p(x))$$

  - We will show two possible approaches to prove this:
    - Approach 1 based on optimization
    - Approach 2 based on KLD properties

- Remark: for simplicity we will use distributions $p(x)$ and $q(x)$ to denote pdfs.

Continuous R.Vs.: max entropy (app 1)

- Recall (we have used it for the uniform distribution)
- We will use the method of Lagrange multipliers (unconstrained regularization)

$$\mathcal{L}(x_1, x_2, ..., x_N) = f(x_1, x_2, ..., x_N) + \lambda\Big(g(x_1, x_2, ..., x_N) - k\Big)$$

$$\left(\hat{x}_1, \hat{x}_2, ..., \hat{x}_N\right) = \underset{x_1, x_2, ..., x_N}{\arg\min}\ \mathcal{L}(x_1, x_2, ..., x_N)$$

- Solution

$$\frac{\partial \mathcal{L}(x_1, x_2, ..., x_N)}{\partial x_i} = \frac{\partial f(x_1, ...x_i, ..., x_N)}{\partial x_i} + \lambda \frac{\partial g(x_1, ...x_i, ..., x_N)}{\partial x_i} = 0$$

$$\frac{d\mathcal{L}(x_1, x_2, ..., x_N)}{d\lambda} = g(x_1, x_2, ..., x_N) - k = 0$$

Continuous R.Vs.: max entropy (app 1)

- We should satisfy the constraints:

$$\int_{-\infty}^{+\infty} q(x)\,dx = 1 \qquad \text{(a)}$$

$$\int_{-\infty}^{+\infty} (x-\mu)^2 q(x)\,dx = \sigma^2 \qquad \text{(b)}$$

- The Lagrangian is:

$$\tilde{\mathcal{L}}\Big(q(x)\Big) = -\int_{-\infty}^{+\infty} q(x)\ln q(x)\,dx + \lambda_0\left(1-\int_{-\infty}^{+\infty} q(x)\,dx\right) + \lambda_1\left(\sigma^2 - \int_{-\infty}^{+\infty}(x-\mu)^2 q(x)\,dx\right)$$

- The maximization of $\tilde{\mathcal{L}}\Big(q(x)\Big)$ is equivalent to minimization of $\mathcal{L}\Big(q(x)\Big) = (-1)\times\tilde{\mathcal{L}}\Big(q(x)\Big)$

$$\mathcal{L}\Big(q(x)\Big) = \int_{-\infty}^{+\infty} q(x)\ln q(x)\,dx - \lambda_0\left(1-\int_{-\infty}^{+\infty} q(x)\,dx\right) - \lambda_1\left(\sigma^2 - \int_{-\infty}^{+\infty}(x-\mu)^2 q(x)\,dx\right)$$

Continuous R.Vs.: max entropy (app 1)

- The derivative wrt $q(x)$ and setting it to zero:

$$\frac{d\mathcal{L}\big(q(x)\big)}{dq(x)} = 1 + \ln q(x) + \lambda_0 + \lambda_1 (x-\mu)^2 = 0$$

$$\ln q(x) = -1 - \lambda_0 - \lambda_1 (x-\mu)^2$$

$$q(x) = e^{-\lambda_0 - 1 - \lambda_1 (x-\mu)^2}$$

  - This is a ”shape” of searched pdf
  - Now we need to find the Lagrangian multipliers by:

$$\frac{d\mathcal{L}\big(q(x)\big)}{d\lambda_0} = 0$$

$$\frac{d\mathcal{L}\big(q(x)\big)}{d\lambda_1} = 0$$

Alternatively we can substitute the found shape to each constraint (a) and (b)

Continuous R.Vs.: max entropy (app 1)

Proof: max entropy for bounded variance

- Satisfy the first constraint:

$$\int q(x)dx = 1 = \int e^{-\lambda_0 - 1 - \lambda_1 (x-\mu)^2} dx$$

$$1 = \int e^{-\lambda_0 - 1 - \lambda_1 z^2} dz \Big| \text{Change of variables } z = x - \mu$$

$$1 = \int e^{-\lambda_0 - 1} e^{-\lambda_1 z^2} dz$$

$$e^{\lambda_0 + 1} = \int e^{-\lambda_1 z^2} dz$$

$$e^{\lambda_0 + 1} = \sqrt{\frac{\pi}{\lambda_1}}$$

Continuous R.Vs.: max entropy (app 1)

- Satisfy the second constraint:

$$\int q(x)(x-\mu)^2 dx = \sigma^2 = \int e^{-\lambda_0 - 1 - \lambda_1 (x-\mu)^2} (x-\mu)^2 dx$$

$$\sigma^2 = \int e^{-\lambda_0 - 1 - \lambda_1 (x-\mu)^2} (x-\mu)^2 dx$$

$$\sigma^2 = \int e^{-\lambda_0 - 1 - \lambda_1 z^2} z^2 dz \Big| \text{Change of variables } z = x - \mu$$

$$\sigma^2 e^{\lambda_0 + 1} = \frac{1}{2}\sqrt{\frac{\pi}{\lambda_1^3}}$$

$$\sigma^2 e^{\lambda_0 + 1} = \frac{1}{2\lambda_1}\sqrt{\frac{\pi}{\lambda_1}}$$

$$2\lambda_1 \sigma^2 e^{\lambda_0 + 1} = \sqrt{\frac{\pi}{\lambda_1}}$$

Continuous R.Vs.: max entropy (app 1)

- Putting together

$$\sqrt{\frac{\pi}{\lambda_1}} = e^{\lambda_0+1} = 2\lambda_1\sigma^2 e^{\lambda_0+1}$$

$$e^{\lambda_0-1} = 2\lambda_1\sigma^2 e^{\lambda_0-1}$$

$$1 = 2\lambda_1\sigma^2$$

$$\lambda_1 = \frac{1}{2\sigma^2}$$

Continuous R.Vs.: max entropy (app 1)

- Putting together

$$\sqrt{\frac{\pi}{\lambda_1}} = e^{\lambda_0+1} = 2\lambda_1\sigma^2 e^{\lambda_0+1}$$

$$e^{\lambda_0-1} = 2\lambda_1\sigma^2 e^{\lambda_0-1}$$

$$1 = 2\lambda_1\sigma^2$$

$$\lambda_1 = \frac{1}{2\sigma^2}$$

$$\sqrt{\frac{\pi}{\lambda_1}} = e^{\lambda_0+1}$$

$$\sqrt{2\sigma^2\pi} = e^{\lambda_0+1}$$

$$\ln\sqrt{2\sigma^2\pi} = \lambda_0 + 1$$

$$\lambda_0 = \ln\sqrt{2\sigma^2\pi} - 1$$

Continuous R.Vs.: max entropy (app 1)

- Substituting $\lambda_1 = \frac{1}{2\sigma^2}$

$$\lambda_0 = \ln\sqrt{2\sigma^2\pi} - 1$$

to $q(x) = e^{-\lambda_0 - 1 - \lambda_1 (x-\mu)^2}$ results

$$q(x) = e^{-\ln\sqrt{2\sigma^2\pi} + 1 - 1 - \frac{1}{2\sigma^2}(x-\mu)^2}$$

$$q(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \Rightarrow \mathcal{N}\left(\mu, \sigma_X^2\right)$$

- Thus, the distribution possessing the maximum entropy among all distributions with the bounded variance is Gaussian.

Continuous R.Vs.: max entropy for fixed var (app 2)

Proof: max entropy for bounded variance

- Gaussian has maximum differential entropy among all distributions with the same variance

- Let $q(x)$ be any density satisfying $\int x^2 q(x)\,dx = \sigma^2$. Let $p(x) = \mathcal{N}\left(0, \sigma_X^2\right)$. Then

$$h(q(x)) \le h(p(x))$$

$$\begin{aligned} 0 \le D_{\mathrm{KL}}(q(x)\|p(x)) &= \int q(x) \log \frac{q(x)}{p(x)} dx \\ &= -h(q(x)) - \int q(x) \log p(x)\, dx \\ &= -h(q(x)) - \int p(x) \log p(x)\, dx \\ &= -h(q(x)) + h(p(x)) \end{aligned}$$

$$h(q(x)) \le h(p(x))$$

We will see all KLD properties later

$$\log p(x) = -\log(\sqrt{2\sigma^2}) - x^2 / 2\sigma^2$$

$$\int q(x) x^2\, dx = \int p(x) x^2\, dx = \sigma^2$$

Continuous R.Vs.: max entropy for fixed var (app 2)

Proof: max entropy for bounded variance

- Demo

$$0 \leq D_{\mathrm{KL}}(q(x)\|p(x)) = \int q(x) \log \frac{q(x)}{p(x)} dx$$

$$-D_{\mathrm{KL}}(q(x)\|p(x)) = \int q(x) \log \frac{p(x)}{q(x)} dx$$

$$\leq \log \int q(x) \frac{p(x)}{q(x)} dx \quad \texttt{(by Jensen's inequality)}$$

$$= \log \int p(x) dx \leq \log = 0$$

Multivariate Gaussian: Differential entropy

Definition (Differential entropy of Gaussian r.v.): *differential entropy of Gaussian random vector* $\mathbf{X} \sim \mathcal{N}(\overline{\mathbf{x}}, \mathbf{K}_{\mathbf{xx}})$ is:

$$h(\mathbf{X}) = -E_{f_{\mathbf{X}}}\left[\log_2 f_{\mathbf{X}}(\mathbf{x})\right] = \frac{1}{2}\ln\left((2\pi e)^N \left|\det \mathbf{K}_{\mathbf{xx}}\right|\right)$$

- Proof

$$f(\mathbf{x}) = f(x_1, x_2, \cdots, x_N) = \frac{1}{\sqrt{(2\pi)^N \left|\det \mathbf{K}_{\mathbf{xx}}\right|}} \exp\left[-\frac{1}{2}(\mathbf{x}-\overline{\mathbf{x}})^T \mathbf{K}_{\mathbf{xx}}{}^{-1}(\mathbf{x}-\overline{\mathbf{x}})\right]$$

$$h(\mathbf{X}) = -\int_{\mathbb{R}^N} f(\mathbf{x}) \ln f(\mathbf{x})\, d\mathbf{x} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(\mathbf{x}) \ln f(\mathbf{x}) d\mathbf{x} =$$

Multivariate Gaussian: Differential entropy

$$
\begin{aligned}
h(\mathbf{X}) &= -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} f(\mathbf{x})\ln f(\mathbf{x})d\mathbf{x} \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} f(\mathbf{x})\left[\frac{1}{2}\ln\left((2\pi)^N\left|\det\mathbf{K}_{\mathbf{xx}}\right|\right)+\frac{1}{2}\left(\mathbf{x}-\overline{\mathbf{x}}\right)^T\mathbf{K}_{\mathbf{xx}}{}^{-1}\left(\mathbf{x}-\overline{\mathbf{x}}\right)\right]d\mathbf{x} \\
&= \frac{1}{2}\ln\left((2\pi)^N\left|\det\mathbf{K}_{\mathbf{xx}}\right|\right)+\frac{1}{2}\underbrace{E_{f(\mathbf{x})}\left[\left(\mathbf{x}-\overline{\mathbf{x}}\right)^T\mathbf{K}_{\mathbf{xx}}{}^{-1}\left(\mathbf{x}-\overline{\mathbf{x}}\right)\right]}_{N}\ln e \\
&= \frac{1}{2}\ln\left((2\pi)^N\left|\det\mathbf{K}_{\mathbf{xx}}\right|\right)+\frac{1}{2}N\ln e \\
&= \frac{1}{2}\ln\left((2\pi e)^N\left|\det\mathbf{K}_{\mathbf{xx}}\right|\right)
\end{aligned}
$$

- We will prove that $E_{f(\mathbf{x})}\left[\left(\mathbf{x}-\overline{\mathbf{x}}\right)^T\mathbf{K}_{\mathbf{xx}}{}^{-1}\left(\mathbf{x}-\overline{\mathbf{x}}\right)\right]=N$

Multivariate Gaussian: Differential entropy

- Note $\underbrace{\left(\mathbf{x}-\overline{\mathbf{x}}\right)^T}_{1\times N}\underbrace{\mathbf{K}_{\mathbf{xx}}{}^{-1}}_{N\times N}\underbrace{\left(\mathbf{x}-\overline{\mathbf{x}}\right)}_{N\times 1}\Rightarrow 1\times 1$

- Since the trace of any 1×1 matrix is $Tr\left[\mathbf{A}\right]=\mathbf{A}$, one obtains

$$\left(\mathbf{x}-\overline{\mathbf{x}}\right)^T\mathbf{K}_{\mathbf{xx}}{}^{-1}\left(\mathbf{x}-\overline{\mathbf{x}}\right)=Tr\left[\left(\mathbf{x}-\overline{\mathbf{x}}\right)^T\mathbf{K}_{\mathbf{xx}}{}^{-1}\left(\mathbf{x}-\overline{\mathbf{x}}\right)\right]$$

- Thus

$$\begin{aligned}E_{f(\mathbf{x})}\left[\left(\mathbf{x}-\overline{\mathbf{x}}\right)^T\mathbf{K}_{\mathbf{xx}}{}^{-1}\left(\mathbf{x}-\overline{\mathbf{x}}\right)\right]&=\int_{-\infty}^{+\infty}f(\mathbf{x})\left(\mathbf{x}-\overline{\mathbf{x}}\right)^T\mathbf{K}_{\mathbf{xx}}{}^{-1}\left(\mathbf{x}-\overline{\mathbf{x}}\right)d\mathbf{x}\\&=\int_{-\infty}^{+\infty}f(\mathbf{x})\ln e^{\left(\mathbf{x}-\overline{\mathbf{x}}\right)^T\mathbf{K}_{\mathbf{xx}}{}^{-1}\left(\mathbf{x}-\overline{\mathbf{x}}\right)}d\mathbf{x}\\&=\int_{-\infty}^{+\infty}f(\mathbf{x})\ln e^{Tr\left[\left(\mathbf{x}-\overline{\mathbf{x}}\right)^T\mathbf{K}_{\mathbf{xx}}{}^{-1}\left(\mathbf{x}-\overline{\mathbf{x}}\right)\right]}d\mathbf{x}\end{aligned}$$

Proof: expectation of quadratic form

Multivariate Gaussian: Differential entropy

- Note $Tr\left[\mathbf{ABC}\right] = Tr\left[\mathbf{CAB}\right] = Tr\left[\mathbf{BCA}\right]$

- Thus

$$\int_{-\infty}^{+\infty} f(\mathbf{x}) \ln e^{\mathrm{Tr}\left[(\mathbf{x}-\overline{\mathbf{x}})^T \mathbf{K}_{\mathbf{xx}}^{-1}(\mathbf{x}-\overline{\mathbf{x}})\right]} d\mathbf{x} = \int_{-\infty}^{+\infty} f(\mathbf{x}) \ln e^{\mathrm{Tr}\left[\mathbf{K}_{\mathbf{xx}}{}^{-1}(\mathbf{x}-\overline{\mathbf{x}})(\mathbf{x}-\overline{\mathbf{x}})^T\right]} d\mathbf{x}$$

- Because $f(\mathbf{x})$ is a scalar and the natural logarithm and exponential may cancel, the properties of the trace function allow us to push $f(\mathbf{x})$ and the integral inside the trace

$$\int_{-\infty}^{+\infty} f(\mathbf{x}) \ln e^{\mathrm{Tr}\left[\mathbf{K}_{\mathbf{xx}}{}^{-1}(\mathbf{x}-\overline{\mathbf{x}})(\mathbf{x}-\overline{\mathbf{x}})^T\right]} d\mathbf{x} = \ln e^{\mathrm{Tr}[\mathbf{K}_{\mathbf{xx}}^{-1} \underbrace{\int_{-\infty}^{+\infty} f(\mathbf{x})(\mathbf{x}-\overline{\mathbf{x}})(\mathbf{x}-\overline{\mathbf{x}})^T d\mathbf{x}}_{\mathbf{K}_{\mathbf{xx}}}]}$$

$$= \ln e^{Tr\left[\mathbf{K}_{\mathbf{xx}}{}^{-1}\mathbf{K}_{\mathbf{xx}}\right]}$$

$$= \ln e^{Tr\left[\mathbf{I}_N\right]} = \ln e^N = N$$

Proof: expectation of quadratic form

Multivariate Gaussian: Differential entropy

Definition (Differential entropy of Gaussian r.v.): *differential entropy of Gaussian random vector* $\mathbf{X} \sim \mathcal{N}(\bar{\mathbf{x}}, \mathbf{K}_{\mathbf{xx}})$ is:

$$h(\mathbf{X}) = \frac{1}{2}\ln\left((2\pi e)^N \det|\mathbf{K}_{\mathbf{xx}}|\right) = \frac{1}{2}\ln\left((2\pi e)^N \prod_{i=1}^{N}\sigma_i\right) = \frac{1}{2}\sum_{i=1}^{N}\ln(2\pi e\sigma_i)$$

$$\mathbf{K}_{\mathbf{xx}} = \mathbf{U}\Sigma\mathbf{U}^T \Rightarrow \left|\det \mathbf{K}_{\mathbf{xx}}\right| = \left|\det \mathbf{U}\Sigma\mathbf{U}^T\right| = \underbrace{\left|\det \mathbf{U}\mathbf{U}^T\right|}_{\mathbf{I}}\left|\det\Sigma\right| = \left|\det\Sigma\right| = \prod_{i=1}^{N}\sigma_i$$

$\Sigma = diag(\sigma_1, \cdots, \sigma_n)$ or $\Sigma_{ii} = \sigma_i, \forall i$ - eigenvalues of covariance matrix (not variances)

$$\mathbf{K}_{x_i x_j} = Cov[X_i, X_j] = \mathrm{E}[(X_i - \mathrm{E}[X_i])(X_j - \mathrm{E}[X_j])] = \sigma_{X_{ij}}$$

$$\mathbf{K}_{x_i x_i} = Cov[X_i, X_i] = \mathrm{E}[(X_i - \mathrm{E}[X_i])(X_i - \mathrm{E}[X_i])] = \sigma^2_{X_i}$$ - variances

Multivariate Gaussian: Differential entropy

Application (Hadamard's inequality) *differential entropy of Gaussian random vector* $\mathbf{X} \sim \mathcal{N}(\bar{\mathbf{x}}, \mathbf{K}_{\mathbf{xx}})$ can be bounded as:

$$h(\mathbf{X}) = \frac{1}{2}\ln\left((2\pi e)^N \left|\det \mathbf{K}_{\mathbf{xx}}\right|\right) \leq \frac{1}{2}\ln\left((2\pi e)^N \prod_{i=1}^{N} \sigma_{X_i}^2\right) = \frac{1}{2}\sum_{i=1}^{N} \ln\left(2\pi e \sigma_{X_i}^2\right)$$

- Hadamard's inequality $\left|\mathbf{K}_{\mathbf{xx}}\right| \leq \prod_{i=1}^{N} \mathbf{K}_{x_i x_i} = \prod_{i=1}^{N} \sigma_{X_i}^2$

$$\mathbf{K}_{\mathbf{xx}} = \begin{bmatrix} \sigma_{X_1}^2 & \rho\sigma_{X_1}\sigma_{X_2} \\ \rho\sigma_{X_1}\sigma_{X_2} & \sigma_{X_2}^2 \end{bmatrix}$$

$$\left|\det \mathbf{K}_{\mathbf{xx}}\right| = \sigma_{X_1}^2\sigma_{X_2}^2 - \underbrace{\rho^2}_{0\leq\rho^2\leq 1}\sigma_{X_1}^2\sigma_{X_2}^2 \leq \sigma_{X_1}^2\sigma_{X_2}^2 = \prod_{i=1}^{N}\sigma_{X_i}^2$$

Multivariate Gaussian: Differential entropy

Application (Hadamard's inequality) *differential entropy of Gaussian random vector* $\mathbf{X} \sim \mathcal{N}(\overline{\mathbf{x}}, \mathbf{K}_{\mathbf{xx}})$ can be bounded as:

$$h\left(\mathbf{X}\right) = \frac{1}{2}\ln\left((2\pi e)^N \left|\det \mathbf{K}_{\mathbf{xx}}\right|\right) \leq \frac{1}{2}\ln\left((2\pi e)^N \prod_{i=1}^{N} \sigma_{X_i}^2\right) = \frac{1}{2}\sum_{i=1}^{N} \ln\left(2\pi e \sigma_{X_i}^2\right)$$

- Alternative "IT proof "

$$\underbrace{\frac{1}{2}\ln\left((2\pi e)^N \left|\det \mathbf{K}_{\mathbf{xx}}\right|\right)}_{\text{general case}} = h\left(\mathbf{X}\right) \underbrace{\leq \sum_{i=1}^{N} h(X_i) =}_{\text{indepenent variables}} \frac{1}{2}\sum_{i=1}^{N} \ln\left(2\pi e \sigma_{X_i}^2\right)$$

Continuous R.Vs.: Differential entropy

- Translation does not change entropy

$$h\left(X+a\right)=h\left(X\right)$$

- Impact of scaling on differential entropy
  - Scalar case

$$h\left(aX\right)=h\left(X\right)+\log\left|a\right|$$

$$X\sim\mathcal{N}\left(0,\sigma_X^2\right)$$

$$aX\sim\mathcal{N}\left(0,a^2\sigma_X^2\right)$$

$$h\left(aX\right)=\frac{1}{2}\log_2\left(2\pi e a^2\sigma_X^2\right)=\frac{1}{2}\log_2\left(2\pi e\sigma_X^2\right)+\log_2\left|a\right|$$

  - Vector case

$$h\left(\mathbf{AX}\right)=h\left(\mathbf{X}\right)+\log\left|\det(\mathbf{A})\right|$$

Continuous R.Vs.: Differential entropy

- Impact of scaling on differential entropy
  - Formal proof

$$h(aX) = h(X) + \log|a|$$

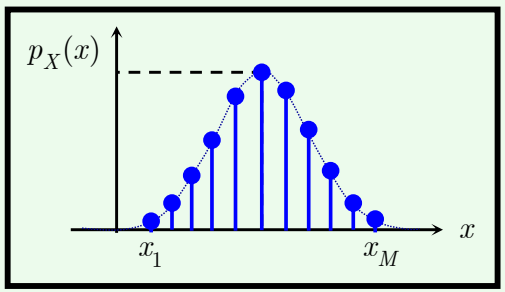
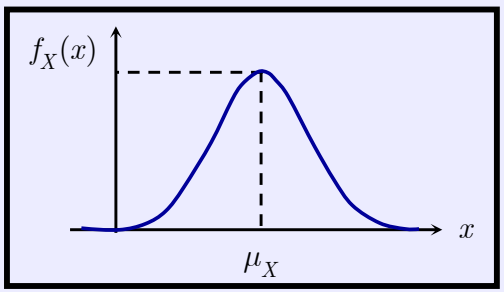
$$Y = aX \qquad f_Y(y) = \frac{1}{|a|} f_X(y/a)$$

$$h(aX) = -\int f_Y(y) \log f_Y(y)\, dy = -\int \frac{1}{|a|} f_X(y/a) \log\left(\frac{1}{|a|} f_X(y/a)\right) dy$$

$$= -\int \frac{1}{|a|} f_X(y/a) \underbrace{\log\left(\frac{1}{|a|}\right)}_{const} dy - \underbrace{\int \frac{1}{|a|} f_X(y/a) \log f_X(y/a)\, dy}_{h(X)}$$

$$= h(X) + \log|a|$$

Summary: entropy of discrete vs continuous RVs

	Entropy of discrete RVs	Entropy of continuous RVs
Negativity	$H(X) \geq 0$	$h(X)$ any
Scaling	$H(aX) = H(X)$	$h(aX) = h(X) + \log\lvert a\rvert$
		

Continuous R.Vs.: max entropy for fixed cov matrix

- Let the random vector $\mathbf{X} \in \mathbb{R}^N$ has zero mean and covariance $\mathbf{K}_{\mathbf{xx}} = E\left[\mathbf{X}\mathbf{X}^T\right]$, i.e., $\mathbf{K}_{x_i x_j} = E\left[X_i X_j\right]$, $1 \le i, j \le N$.

- Then any random vector with the above covariance matrix has the entropy smaller or equal to the Gaussian one $h(\mathbf{X}) \le \frac{1}{2}\ln\left((2\pi e)^N \left|\mathbf{K}_{\mathbf{xx}}\right|\right)$

  - Let $q(\mathbf{x})$ be any density satisfying $\int \mathbf{x}_i \mathbf{x}_j q(\mathbf{x}) d\mathbf{x} = E_q\left[X_i X_j\right] = \mathbf{K}_{x_i x_j}$

  - Let $p(\mathbf{x})$ be the density of a $\mathcal{N}\left(\mathbf{0}, \mathbf{K}_{\mathbf{xx}}\right)$, i.e., it has $\int \mathbf{x}_i \mathbf{x}_j p(\mathbf{x}) d\mathbf{x} = \mathbf{K}_{x_i x_j}$

    - Note that $\log p(\mathbf{x}) = -\log\left(\sqrt{\left(2\pi\right)^N \left|\det \mathbf{K}_{\mathbf{xx}}\right|}\right) - \left[-\frac{1}{2}\mathbf{x}^T \mathbf{K}_{\mathbf{xx}}{}^{-1}\mathbf{x}\right]$

Continuous R.Vs.: max entropy for fixed cov matrix

$$0 \leq D_{\mathrm{KL}}(q(\mathbf{x})\|p(\mathbf{x})) = \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x}$$

$$= -h(q(\mathbf{x})) - \int q(\mathbf{x}) \log p(\mathbf{x})\, d\mathbf{x}$$

$$= -h(q(\mathbf{x})) - \int p(\mathbf{x}) \log p(\mathbf{x})\, d\mathbf{x}$$

$$= -h(q(\mathbf{x})) + h(p(\mathbf{x}))$$

$$h(q(\mathbf{x})) \leq h(p(\mathbf{x}))$$

$$\int q(\mathbf{x}) \log p(\mathbf{x})\, d\mathbf{x} = \int p(\mathbf{x}) \log p(\mathbf{x})\, d\mathbf{x}$$

$$\log p(\mathbf{x}) = -\log\left(\sqrt{(2\pi)^N \left|\det \mathbf{K}_{\mathbf{xx}}\right|}\right) - \left[-\frac{1}{2}\mathbf{x}^T \mathbf{K}_{\mathbf{xx}}{}^{-1}\mathbf{x}\right]$$

$$\int q(\mathbf{x})\mathbf{x}^T \mathbf{K}_{\mathbf{xx}}{}^{-1}\mathbf{x} d\mathbf{x} = N$$

$$\int p(\mathbf{x})\mathbf{x}^T \mathbf{K}_{\mathbf{xx}}{}^{-1}\mathbf{x} d\mathbf{x} = N$$

See “Proof of expectation of quadratic form”

Continuous R.V.: relative entropy

Definition (relative entropy): *Relative entropy* or *Kullback-Leibler divergence* between pdfs $p(x)$ and $q(x)$:

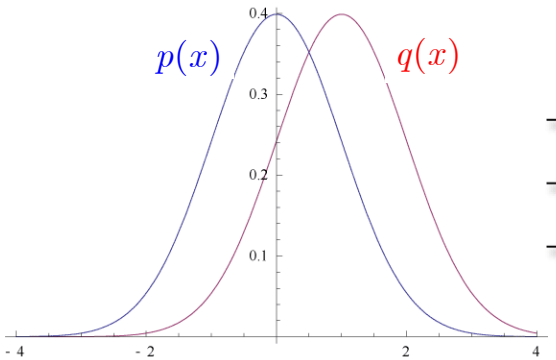
$$D_{\mathrm{KL}}(p \parallel q) = \int_{-\infty}^{\infty} p(x) \log \left(\frac{p(x)}{q(x)} \right) dx = E_{p(x)} \left[\log_2 \frac{p(x)}{q(x)} \right]$$

- Equivalent notations

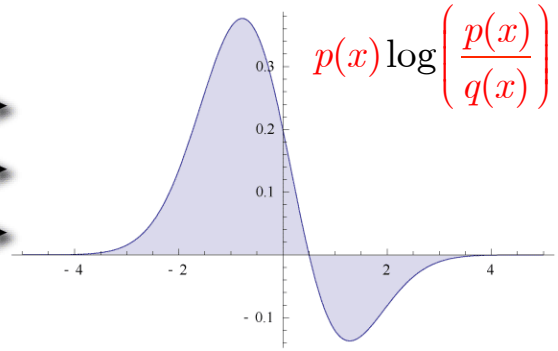
$$D_{\mathrm{KL}}(f \| q) = \int_{\mathcal{X}} f_X(x) \log_2 \frac{f_X(x)}{q_X(x)} dx$$

Continuous R.V.: relative entropy

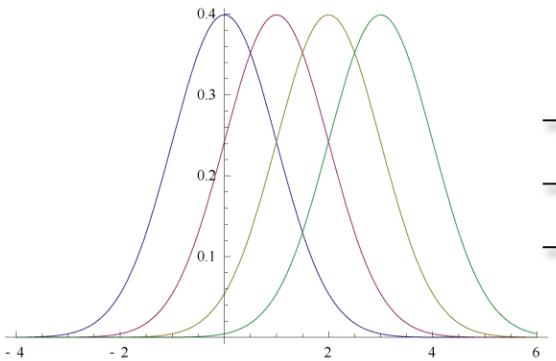
$$D_{\mathrm{KL}}(p \parallel q) = \int_{-\infty}^{\infty} p(x) \log \left(\frac{p(x)}{q(x)} \right) dx = \underbrace{E_{p(x)}\left[\log_2 p\left(x\right)\right]}_{-h(p(x))} - \underbrace{E_{p(x)}\left[\log_2 q\left(x\right)\right]}_{h(p(x);q(x))}$$

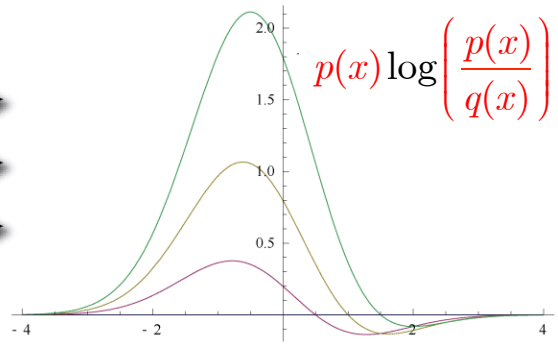


Original Gaussian PDF's



KL Area to be Integrated $D_{\mathrm{KL}}(p \parallel q)$





Continuous R.Vs.: KLD between two Gaussians

Definition (relative entropy between two Gaussian): *Relative entropy* between two univaraite Gaussian pdfs $p(x) = \mathcal{N}(\mu_p, \sigma_p^2)$ and $q(x) = \mathcal{N}(\mu_q, \sigma_q^2)$

$$D_{\mathrm{KL}}(p \parallel q) = D_{\mathrm{KL}}(\mathcal{N}(\mu_p, \sigma_p^2) \parallel \mathcal{N}(\mu_q, \sigma_q^2))$$
$$= \log \sigma_q - \log \sigma_p - \frac{1}{2}\left[1 - \left(\frac{\sigma_p^2 + \left(\mu_p - \mu_q\right)^2}{\sigma_q^2}\right)\right]$$

- Proof

$$D_{\mathrm{KL}}(p \parallel q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx = E_{p(x)}\left[\log_2 \frac{p\left(x\right)}{q\left(x\right)}\right]$$

Continuous R.Vs.: KLD between two Gaussians

Proof: KLD between two univariate Gaussians

$$\log \frac{p(x)}{q(x)} = \log \left(\frac{\frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left[-\frac{1}{2\sigma_p^2} \left(x - \mu_p \right)^2 \right]}{\frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left[-\frac{1}{2\sigma_q^2} \left(x - \mu_q \right)^2 \right]} \right)$$

$$= \log \left(\frac{\sqrt{2\pi\sigma_q^2}}{\sqrt{2\pi\sigma_p^2}} \right) + \frac{1}{2\sigma_q^2} \left(x - \mu_q \right)^2 - \frac{1}{2\sigma_p^2} \left(x - \mu_p \right)^2$$

$$D_{\mathrm{KL}}(p \parallel q) = E_{p(x)} \left[\log_2 \frac{p(x)}{q(x)} \right] = \frac{1}{2} \log \left(\frac{\sigma_q^2}{\sigma_p^2} \right) + \frac{1}{2\sigma_q^2} E_{p(x)} \left[\left(X - \mu_q \right)^2 \right]$$

$$- \frac{1}{2\sigma_p^2} \underbrace{E_{p(x)} \left[\left(X - \mu_p \right)^2 \right]}_{\sigma_p^2}$$

Continuous R.Vs.: KLD between two Gaussians

Proof: KLD between two univariate Gaussians

$$E_{p(x)}\left[\left(X-\mu_q\right)^2\right] = E_{p(x)}\left[X^2 - 2X\mu_q + \mu_q^2\right]$$

$$= \underbrace{E_{p(x)}\left[X^2\right]}_{\sigma_p^2+\mu_p^2} - 2\mu_q \underbrace{E_{p(x)}\left[X\right]}_{\mu_p} + \mu_q^2$$

$$= \sigma_p^2 + \mu_p^2 - 2\mu_q\mu_p + \mu_q^2$$

$$= \sigma_p^2 - (\mu_p - \mu_q)^2$$

$$D_{\mathrm{KL}}(p \parallel q) = \frac{1}{2}\log\left(\frac{\sigma_q^2}{\sigma_p^2}\right) + \frac{1}{2\sigma_q^2}E_{p(x)}\left[\left(X-\mu_q\right)^2\right] - \frac{1}{2\sigma_p^2}E_{p(x)}\left[\left(X-\mu_p\right)^2\right]$$

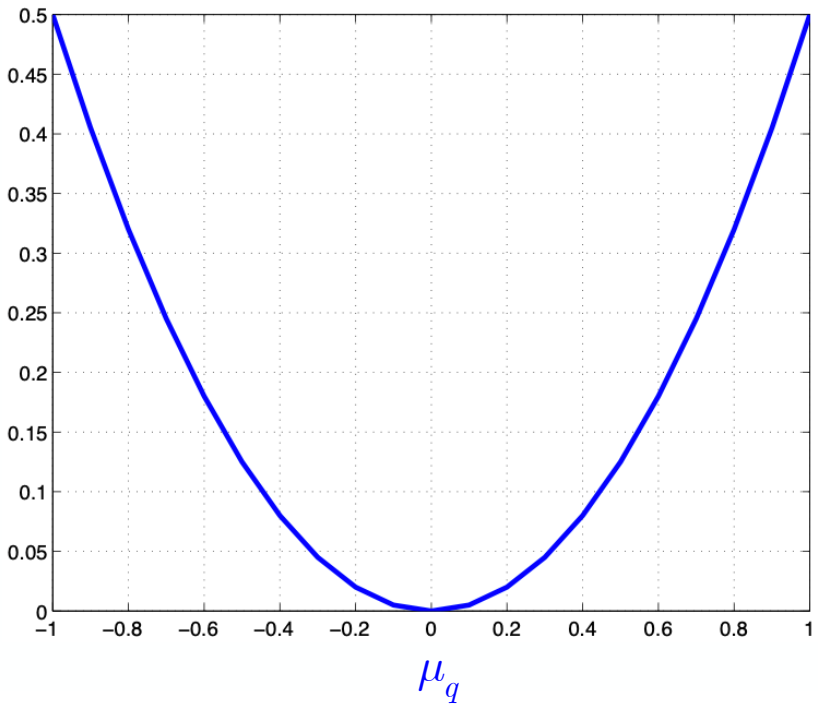
$$= \log\sigma_q - \log\sigma_p + \frac{1}{2\sigma_q^2}\left[\sigma_p^2 - (\mu_p-\mu_q)^2\right] - \frac{1}{2\sigma_p^2}\sigma_p^2$$

$$= \log\sigma_q - \log\sigma_p - \frac{1}{2}\left[1 - \left(\frac{\sigma_p^2 + \left(\mu_p-\mu_q\right)^2}{\sigma_q^2}\right)\right]$$

Continuous R.Vs.: KLD between two Gaussians

$$p(x) = \mathcal{N}(0,1) \qquad q(x) = \mathcal{N}(\mu_q, 1)$$

$D_{\mathrm{KL}}(p \parallel q)$



Continuous R.V.: miltivariate relative entropy

Definition (relative entropy): *Relative entropy* or *Kullback-Leibler divergence* between pdfs $p\left(\mathbf{x}\right)$ and $q\left(\mathbf{x}\right)$:

$$D_{\mathrm{KL}}(p \parallel q) = \int_{\mathbb{R}^N} p(\mathbf{x}) \log \left(\frac{p(\mathbf{x})}{q(\mathbf{x})} \right) d\mathbf{x} = E_{p(\mathbf{x})} \left[\log_2 \frac{p\left(\mathbf{x}\right)}{q\left(\mathbf{x}\right)} \right]$$

Continuous R.V.: KLD for multivariate Gaussians

Definition (relative entropy): *Kullback-Leibler divergence* between two multivariate Gaussian pdfs $p(\mathbf{x}) = \mathcal{N}(\overline{\mathbf{x}}_p, \mathbf{K}_p)$ and $q(\mathbf{x}) = \mathcal{N}(\overline{\mathbf{x}}_q, \mathbf{K}_q)$:

$$D_{\mathrm{KL}}\left(p(\mathbf{x}) \parallel q(\mathbf{x})\right) = \frac{1}{2}\left(\ln\left(\frac{\det \mathbf{K}_q}{\det \mathbf{K}_p}\right) + \mathrm{tr}\left(\mathbf{K}_q^{-1}\mathbf{K}_p\right) + \left(\overline{\mathbf{x}}_q - \overline{\mathbf{x}}_p\right)^T \mathbf{K}_q^{-1}\left(\overline{\mathbf{x}}_q - \overline{\mathbf{x}}_p\right) - N\right)$$

- Proof: we will consider two approaches to prove it
  - Approach A: direct proof
  - Approach B: proof via entropy-cross entropy decomposition

Continuous R.V.: KLD for multivariate Gaussians

- Recall $$p(\mathbf{x}) = \frac{1}{\sqrt{\left(2\pi\right)^N \left|\det \mathbf{K}_p\right|}} \exp\left[-\frac{1}{2}\left(\mathbf{x}-\overline{\mathbf{x}}_p\right)^T \mathbf{K}_p^{-1}\left(\mathbf{x}-\overline{\mathbf{x}}_p\right)\right]$$

$$\ln p(\mathbf{x}) = -\frac{1}{2}\ln\left(\left(2\pi\right)^N \left|\det \mathbf{K}_p\right|\right) - \frac{1}{2}\left(\mathbf{x}-\overline{\mathbf{x}}_p\right)^T \mathbf{K}_p^{-1}\left(\mathbf{x}-\overline{\mathbf{x}}_p\right)$$

- Substitute $p(\mathbf{x}) = \mathcal{N}(\overline{\mathbf{x}}_p, \mathbf{K}_p)$ and $q(\mathbf{x}) = \mathcal{N}(\overline{\mathbf{x}}_q, \mathbf{K}_q)$

$$\begin{aligned}
D_{\mathrm{KL}}\left(p(\mathbf{x}) \parallel q(\mathbf{x})\right) &= E_{p(\mathbf{x})}\left[\ln p(\mathbf{x}) - \ln q(\mathbf{x})\right] \\
&= \frac{1}{2}\mathrm{E}_{p(\mathbf{x})}\left[-\ln\left(\left|\det \mathbf{K}_p\right|\right) - \left(\mathbf{x}-\overline{\mathbf{x}}_p\right)^T \mathbf{K}_p^{-1}\left(\mathbf{x}-\overline{\mathbf{x}}_p\right) + \ln\left(\left|\det \mathbf{K}_q\right|\right) + \left(\mathbf{x}-\overline{\mathbf{x}}_q\right)^T \mathbf{K}_q^{-1}\left(\mathbf{x}-\overline{\mathbf{x}}_q\right)\right] \\
&= \frac{1}{2}\ln\left(\frac{\left|\det \mathbf{K}_q\right|}{\left|\det \mathbf{K}_p\right|}\right) + \frac{1}{2}\mathrm{E}_{p(\mathbf{x})}\left[-\left(\mathbf{x}-\overline{\mathbf{x}}_p\right)^T \mathbf{K}_p^{-1}\left(\mathbf{x}-\overline{\mathbf{x}}_p\right) + \left(\mathbf{x}-\overline{\mathbf{x}}_q\right)^T \mathbf{K}_q^{-1}\left(\mathbf{x}-\overline{\mathbf{x}}_q\right)\right] \\
&= \frac{1}{2}\ln\left(\frac{\left|\det \mathbf{K}_q\right|}{\left|\det \mathbf{K}_p\right|}\right) + \frac{1}{2}\mathrm{E}_{p(\mathbf{x})}\left[-Tr\left[\mathbf{K}_p^{-1}\left(\mathbf{x}-\overline{\mathbf{x}}_p\right)\left(\mathbf{x}-\overline{\mathbf{x}}_p\right)^T\right] + Tr\left[\mathbf{K}_q^{-1}\left(\mathbf{x}-\overline{\mathbf{x}}_q\right)\left(\mathbf{x}-\overline{\mathbf{x}}_q\right)^T\right]\right]
\end{aligned}$$

Continuous R.V.: KLD for multivariate Gaussians

$$\ldots = \frac{1}{2}\ln\left(\frac{|\det \mathbf{K}_q|}{|\det \mathbf{K}_p|}\right) + \frac{1}{2}\mathrm{E}_{p(\mathbf{x})}\left[-Tr\left[\mathbf{K}_p^{-1}\left(\mathbf{x}-\overline{\mathbf{x}}_p\right)\left(\mathbf{x}-\overline{\mathbf{x}}_p\right)^T\right] + Tr\left[\mathbf{K}_q^{-1}\left(\mathbf{x}-\overline{\mathbf{x}}_q\right)\left(\mathbf{x}-\overline{\mathbf{x}}_q\right)^T\right]\right]$$

$$\mathrm{E}_{p(\mathbf{x})}\left[-Tr\left[\mathbf{K}_p^{-1}\left(\mathbf{x}-\overline{\mathbf{x}}_p\right)\left(\mathbf{x}-\overline{\mathbf{x}}_p\right)^T\right]\right] = -Tr\left[\mathbf{K}_p^{-1}\mathrm{E}_{p(\mathbf{x})}\left[\left(\mathbf{x}-\overline{\mathbf{x}}_p\right)\left(\mathbf{x}-\overline{\mathbf{x}}_p\right)^T\right]\right]$$
$$= -Tr\left[\mathbf{K}_p^{-1}\mathbf{K}_p\right] = -Tr\left[\mathbf{I}_N\right] = -N$$

$$\mathrm{E}_{p(\mathbf{x})}\left[Tr\left[\mathbf{K}_q^{-1}\left(\mathbf{x}-\overline{\mathbf{x}}_q\right)\left(\mathbf{x}-\overline{\mathbf{x}}_q\right)^T\right]\right] = \mathrm{E}_{p(\mathbf{x})}\left[Tr\left[\mathbf{K}_q^{-1}\left[\mathbf{x}\mathbf{x}^T - 2\mathbf{x}\overline{\mathbf{x}}_q^T + \overline{\mathbf{x}}_q\overline{\mathbf{x}}_q^T\right]\right]\right]$$
$$= Tr\left[\mathbf{K}_q^{-1}\left[\underbrace{\mathrm{E}_{p(\mathbf{x})}\left[\mathbf{x}\mathbf{x}^T\right]}_{\mathbf{K}_p+\overline{\mathbf{x}}_p\overline{\mathbf{x}}_p^T} - 2\underbrace{\mathrm{E}_{p(\mathbf{x})}\left[\mathbf{x}\right]}_{\overline{\mathbf{x}}_p}\overline{\mathbf{x}}_q^T + \overline{\mathbf{x}}_q\overline{\mathbf{x}}_q^T\right]\right]$$
$$= Tr\left[\mathbf{K}_q^{-1}\left[\mathbf{K}_p + \overline{\mathbf{x}}_p\overline{\mathbf{x}}_p^T - 2\overline{\mathbf{x}}_p\overline{\mathbf{x}}_q^T + \overline{\mathbf{x}}_q\overline{\mathbf{x}}_q^T\right]\right]$$
$$= Tr\left[\mathbf{K}_q^{-1}\left[\mathbf{K}_p + \overline{\mathbf{x}}_p\overline{\mathbf{x}}_p^T - 2\overline{\mathbf{x}}_p\overline{\mathbf{x}}_q^T + \overline{\mathbf{x}}_q\overline{\mathbf{x}}_q^T\right]\right]$$
$$= Tr\left[\mathbf{K}_q^{-1}\mathbf{K}_p\right] + Tr\left[\mathbf{K}_q^{-1}\left[\overline{\mathbf{x}}_p\overline{\mathbf{x}}_p^T - 2\overline{\mathbf{x}}_p\overline{\mathbf{x}}_q^T + \overline{\mathbf{x}}_q\overline{\mathbf{x}}_q^T\right]\right]$$
$$= Tr\left[\mathbf{K}_q^{-1}\mathbf{K}_p\right] + Tr\left[\underbrace{\overline{\mathbf{x}}_p\mathbf{K}_q^{-1}\overline{\mathbf{x}}_p^T - 2\overline{\mathbf{x}}_p\mathbf{K}_q^{-1}\overline{\mathbf{x}}_q^T + \overline{\mathbf{x}}_q\mathbf{K}_q^{-1}\overline{\mathbf{x}}_q^T}_{\left(\overline{\mathbf{x}}_q-\overline{\mathbf{x}}_p\right)^T\mathbf{K}_q^{-1}\left(\overline{\mathbf{x}}_q-\overline{\mathbf{x}}_p\right)}\right]$$
$$= Tr\left[\mathbf{K}_q^{-1}\mathbf{K}_p\right] + \left(\overline{\mathbf{x}}_q-\overline{\mathbf{x}}_p\right)^T\mathbf{K}_q^{-1}\left(\overline{\mathbf{x}}_q-\overline{\mathbf{x}}_p\right)$$

Continuous R.V.: KLD for multivariate Gaussians

- Recall $D_{\mathrm{KL}}(p \parallel q) = h(p,q) - h(p)$

$$h(p) = -E_{p(\mathbf{x})}\left[\ln p(\mathbf{x})\right] = \frac{1}{2}\ln\left((2\pi e)^N \det\left|\mathbf{K}_p\right|\right) = \frac{1}{2}\ln\left((2\pi)^N \det\left|\mathbf{K}_p\right|\right) + \frac{1}{2}\underbrace{\ln\left(e^N\right)}_{N}$$

$$h(p,q) = -E_{p(\mathbf{x})}\left[\ln q(\mathbf{x})\right]$$

$$\ln q(\mathbf{x}) = -\frac{1}{2}\ln\left(\left(2\pi\right)^N\left|\det\mathbf{K}_q\right|\right) - \frac{1}{2}\left(\mathbf{x}-\bar{\mathbf{x}}_q\right)^T\mathbf{K}_q^{-1}\left(\mathbf{x}-\bar{\mathbf{x}}_q\right)$$

$$h(p,q) = -E_{p(\mathbf{x})}\left[-\frac{1}{2}\ln\left(\left(2\pi\right)^N\left|\det\mathbf{K}_q\right|\right) - \frac{1}{2}\left(\mathbf{x}-\bar{\mathbf{x}}_q\right)^T\mathbf{K}_q^{-1}\left(\mathbf{x}-\bar{\mathbf{x}}_q\right)\right]$$

$$= \frac{1}{2}\ln\left(\left(2\pi\right)^N\left|\det\mathbf{K}_q\right|\right) + \underbrace{E_{p(\mathbf{x})}\left[\frac{1}{2}\left(\mathbf{x}-\bar{\mathbf{x}}_q\right)^T\mathbf{K}_q^{-1}\left(\mathbf{x}-\bar{\mathbf{x}}_q\right)\right]}_{Tr\left[\mathbf{K}_q^{-1}\mathbf{K}_p\right]+\left(\bar{\mathbf{x}}_q-\bar{\mathbf{x}}_p\right)^T\mathbf{K}_q^{-1}\left(\bar{\mathbf{x}}_q-\bar{\mathbf{x}}_p\right)}$$

Continuous R.V.: miltivariate relative entropy

Relative entropy between $p(\mathbf{x}) = \mathcal{N}(\overline{\mathbf{x}}_p, \Sigma_p)$ and $q(\mathbf{x}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$

$$D_{\mathrm{KL}}\left(\mathcal{N}\left(\left(\mu_1,\ldots,\mu_N\right)^T, \mathrm{diag}\left(\sigma_1^2,\ldots,\sigma_N^2\right)\right) \parallel \mathcal{N}(\mathbf{0},\mathbf{I})\right) = \frac{1}{2}\sum_{i=1}^{N}\left(\sigma_i^2 + \mu_i^2 - \ln\left(\sigma_i^2\right) - 1\right)$$

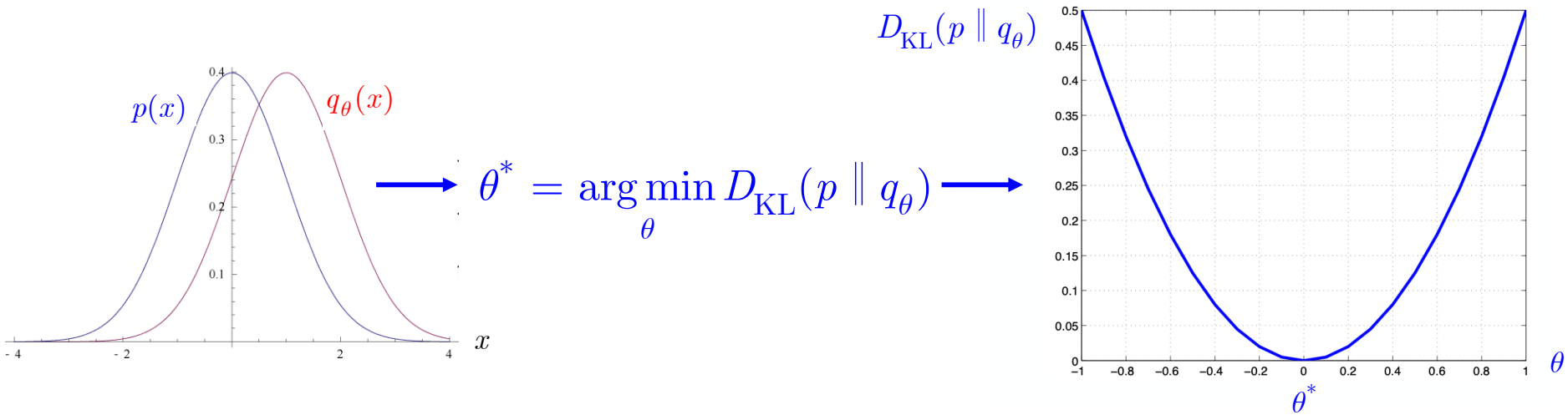
$$\overline{\mathbf{x}}_p = \begin{bmatrix}\mu_1 \\ \vdots \\ \mu_N\end{bmatrix} \quad \Sigma_p = \begin{bmatrix}\sigma_1^2 & & \\ & \ddots & \\ & & \sigma_N^2\end{bmatrix} \qquad \overline{\mathbf{x}}_q = \begin{bmatrix}0 \\ \vdots \\ 0\end{bmatrix} \quad \Sigma_q = \begin{bmatrix}1 & & \\ & \ddots & \\ & & 1\end{bmatrix}$$

KLD in action

- We consider properties of KLD on several ML applications
  - Approximation of true pdf by approximate pdf
  - Binary hypothesis testing

Continuous R.V.: KLD for approximation

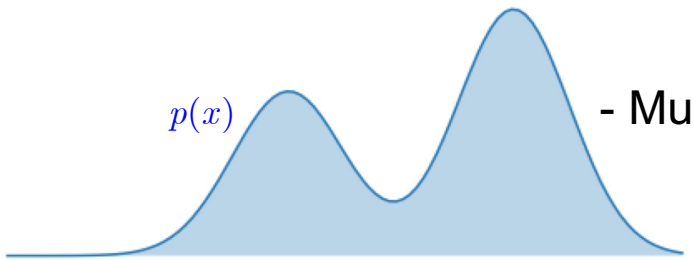
- We consider properties of KLD on one major ML application
- Approximation of a given true distribution $p(\mathbf{x})$ by some approximative distribution $q_\theta(\mathbf{x})$ controlled by a set of parameters θ



- KLD is used as a "measure" of closeness of approximation
- Our problem: should we use $D_{\mathrm{KL}}(p \parallel q_\theta)$ or $D_{\mathrm{KL}}(q_\theta \parallel p)$?
  - Recall $D_{\mathrm{KL}}(p \parallel q_\theta) \neq D_{\mathrm{KL}}(q_\theta \parallel p)$

Continuous R.V.: Forward and Reverse KLD

- We have two options:
  - Forward KLD: $\arg\min_{\theta} D_{\mathrm{KL}}(p \parallel q_\theta)$
  - Reverse KLD: $\arg\min_{\theta} D_{\mathrm{KL}}(q_\theta \parallel p)$
- Example: let $q(x) = \mathcal{N}(\mu, \sigma^2); \theta = \left(\mu, \sigma^2\right)$ - unimodal (simple) pdf

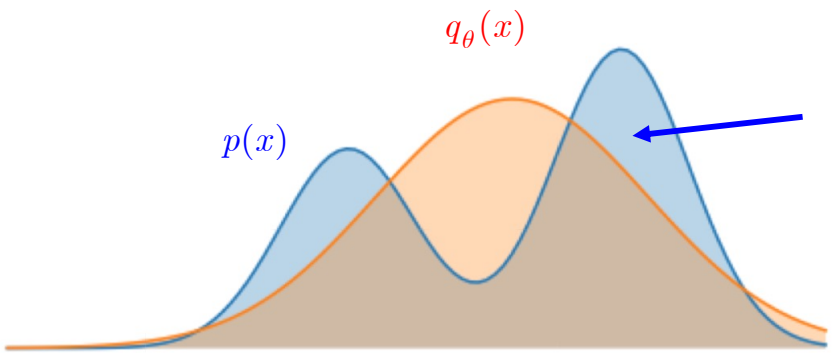


- Multimodal and asymmetric (complex) pdf

Continuous R.V.: Forward and Reverse KLD

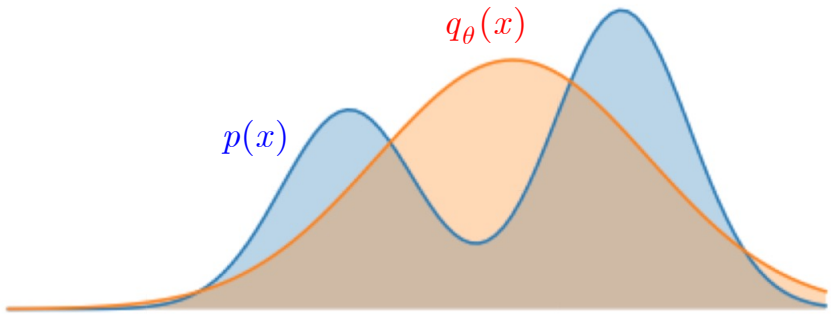
- **Forward KL: Mean-Seeking Behavior**

$$\begin{aligned}
\arg\min_\theta D_{\mathrm{KL}}(p \parallel q_\theta) &= \arg\min_\theta \underbrace{E_{p(x)}\big[\log p(x)\big]}_{-H(p(x)) \text{ const wrt } \theta} \underbrace{-E_{p(x)}\big[\log q_\theta(x)\big]}_{+H(p(x);q_\theta(x))} \\
&\equiv \arg\min_\theta \underbrace{-E_{p(x)}\big[\log q_\theta(x)\big]}_{H(p(x);q_\theta(x))} \\
&= \arg\max_\theta \underbrace{E_{p(x)}\big[\log q_\theta(x)\big]}_{H(p(x);q_\theta(x))} \\
&= \arg\min_\theta H(p(x);q_\theta(x))
\end{aligned}$$



Continuous R.V.: Forward and Reverse KLD

- **Forward KL: Mean-Seeking Behavior**



  - The approximate distribution centers itself between the two modes, so that it can have high coverage of both

  - The maximum (mode) of approximate distribution does not coincide with the mode of true distribution

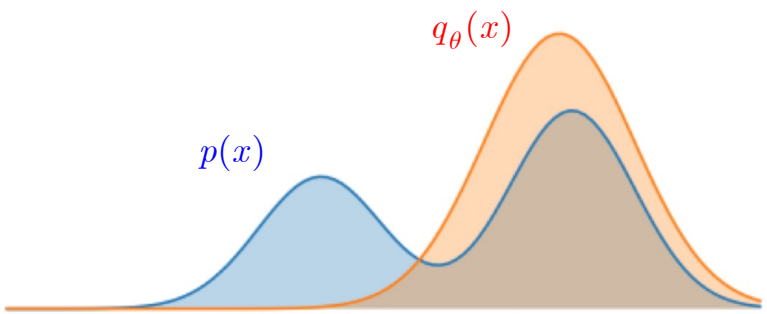
Continuous R.V.: Forward and Reverse KLD

- **Reverse KL: Mode-Seeking Behavior**

$$\arg\min_\theta D_{\mathrm{KL}}(q_\theta \parallel p) = \arg\min_\theta \underbrace{E_{q_\theta(x)}\big[\log q_\theta(x)\big]}_{-H(q_\theta(x))} \underbrace{-E_{q_\theta(x)}\big[\log p(x)\big]}_{+H(q_\theta(x);p(x))}$$

$$\equiv \arg\min_\theta - H(q_\theta(x)) + H(q_\theta(x); p(x))$$

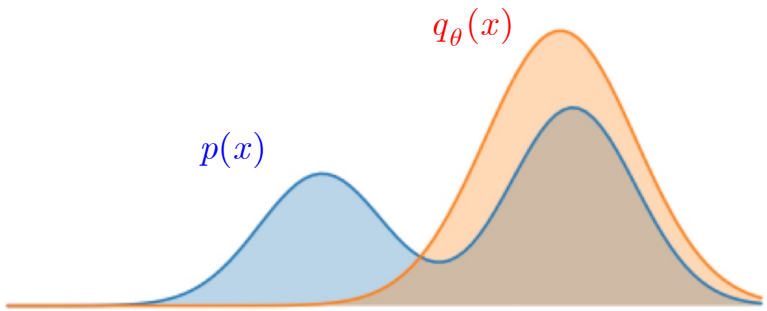
$$= \arg\max_\theta H(q_\theta(x)) - H(q_\theta(x); p(x))$$



The best fit with the regularization that the searched distribution should have max entropy at the same time

Continuous R.V.: Forward and Reverse KLD

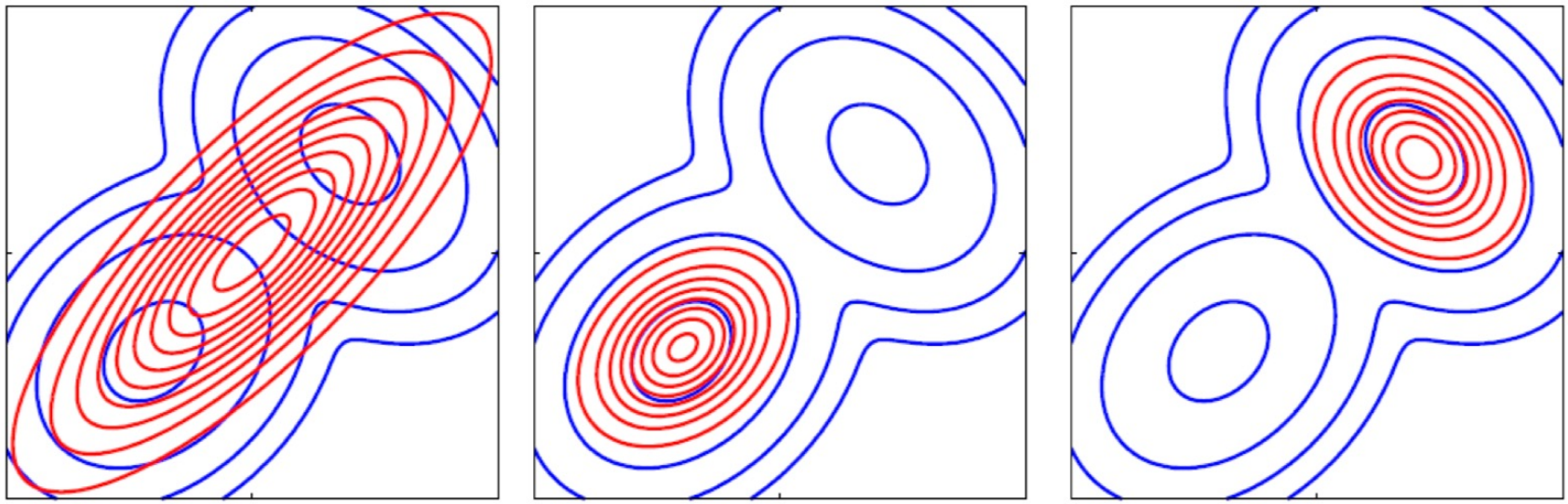
- **Reverse KL: Mode-Seeking Behavior**



- The approximate distribution centers itself at the mode
- At the same time, the cross-entropy term prevents the approximate distribution from collapsing to a very narrow mode
- Thus, optimizing this objective is to find a mode of true distribution with high probability and wide support and mimic true distribution

Continuous R.V.: Forward and Reverse KLD

- **Forward and reverse KLD in 2D**



$$\arg\min_{\theta} D_{\mathrm{KL}}(p \parallel q_\theta)$$

$$\arg\min_{\theta} D_{\mathrm{KL}}(q_\theta \parallel p)$$

Figure from *Bishop, Pattern Recognition and Machine Learning, 2006*

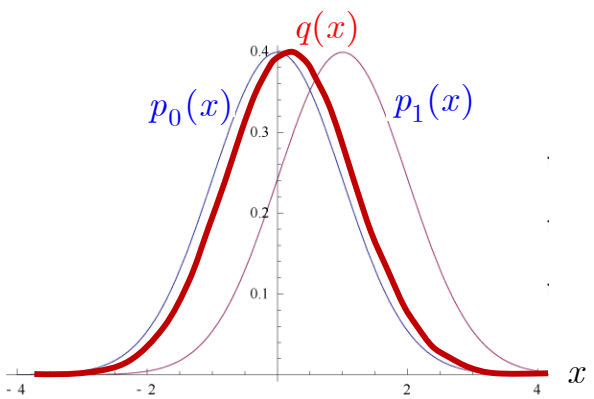
KLD in action

- We consider properties of KLD on several ML applications
  - Approximation of true pdf by approximate pdf
  - Binary hypothesis testing (HT)

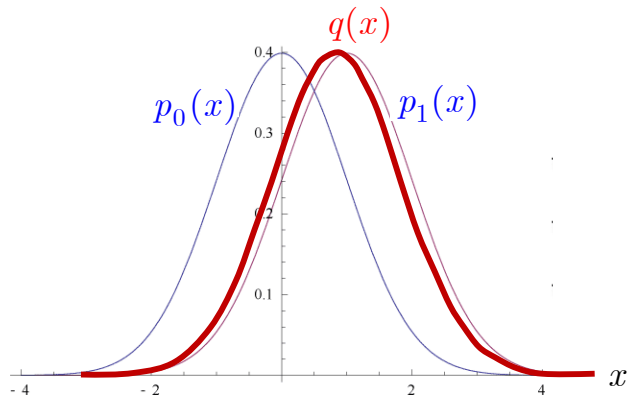
Continuous R.V.: KLD for binary HT

- We consider the use of KLD for binary sequential hypothesis testing

- Given a sequence of observation $X_1, X_2, \cdots, X_N \overset{i.i.d.}{\sim} q(x)$
- Given two models $H_0 : p_0(x)$ and $H_1 : p_1(x)$, we should decide from which distribution the above sequence is generated from or "belongs" to
  - Intuition: what is the best fit $D_{\mathrm{KL}}(q \parallel p_0)$ or $D_{\mathrm{KL}}(q \parallel p_1)$?



$D_{\mathrm{KL}}(q \parallel p_0) < D_{\mathrm{KL}}(q \parallel p_1) \Rightarrow H_0$



$D_{\mathrm{KL}}(q \parallel p_0) > D_{\mathrm{KL}}(q \parallel p_1) \Rightarrow H_1$

Continuous R.V.: KLD for binary HT

- If $q(x)$ is know, we know how to do it
- However, we only know samples $X_1, X_2, \cdots, X_N \overset{i.i.d.}{\sim} q(x)$ generated from $q(x)$

- Consider likelihood ratio test (LRT) for N independent observations

$$\mathcal{L}(X_1, X_2, \cdots, X_N) = \prod_{i=1}^{N} \frac{p_1(x_i)}{p_0(x_i)} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma$$

- Normalized (divided by N) log-likelihood ratio test

$$\ell(X_1, X_2, \cdots, X_N) = \frac{1}{N} \log \mathcal{L}(X_1, X_2, \cdots, X_N) = \frac{1}{N} \sum_{i=1}^{N} \log \frac{p_1(x_i)}{p_0(x_i)} \underset{H_0}{\overset{H_1}{\gtrless}} \ln \gamma$$

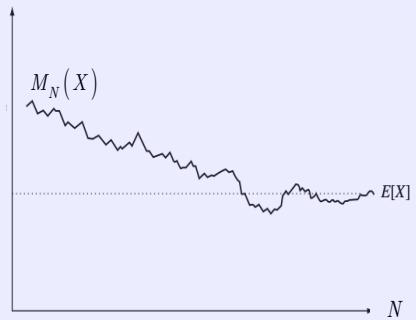
Continuous R.V.: KLD for binary HT

- Remark: $\ell(X_1, X_2, \cdots, X_N)$ is a random variable representing a sum
- From the strong low of large numbers

  - Let $X_1, X_2, \cdots, X_N$ be a sequence of i.i.d. random variables with finite mean $E[X_i] = \mu_X$ and variance $Var[X_i] = \sigma_X^2$ (some bounded)

  - In this case

$$\lim_{N\to\infty} \Pr\left[\lim_{N\to\infty} M_N(X) = \mu_X\right] = 1$$



  - $M_K(X)$ is the sample mean computed using K variables

$$\frac{1}{N}\sum_{i=1}^{N} \log \frac{p_1(x_i)}{p_0(x_i)} \overset{\text{as } N\to\infty}{\to} E_{q(x)}\left[\log \frac{p_1(x)}{p_0(x)}\right]$$

Continuous R.V.: KLD for binary HT

$$
\begin{aligned}
E_{q(x)}\left[\log\frac{p_1(x)}{p_0(x)}\right] &= E_{q(x)}\left[\log\frac{p_1(x)}{p_0(x)}\frac{q(x)}{q(x)}\right] \\
&= E_{q(x)}\left[\log\frac{q(x)}{p_0(x)}\right] - E_{q(x)}\left[\log\frac{q(x)}{p_1(x)}\right] \\
&= D_{\mathrm{KL}}(q \parallel p_0) - D_{\mathrm{KL}}(q \parallel p_1)
\end{aligned}
$$

$$
\frac{1}{N}\sum_{i=1}^{N}\log\frac{p_1(x_i)}{p_0(x_i)} \underset{H_0}{\overset{H_1}{\gtrless}} \ln\gamma \quad \Rightarrow \quad D_{\mathrm{KL}}(q \parallel p_0) - D_{\mathrm{KL}}(q \parallel p_1) \underset{H_0}{\overset{H_1}{\gtrless}} \ln\gamma
$$

- Remark: for equal probabilities of hypothesis and minimum of average probability of errors $\ln\gamma = 0$:

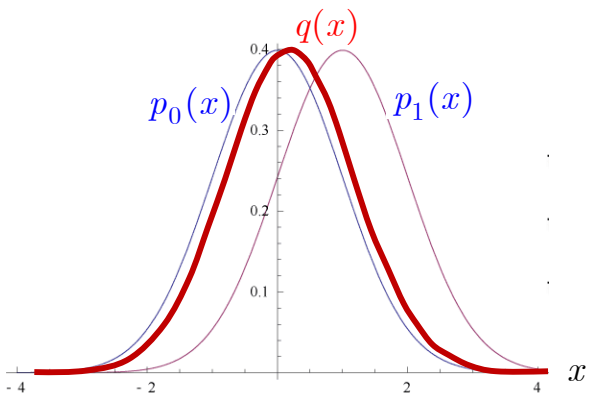
$$
D_{\mathrm{KL}}(q \parallel p_0) \underset{H_0}{\overset{H_1}{\gtrless}} D_{\mathrm{KL}}(q \parallel p_1)
$$

Continuous R.V.: KLD for binary HT

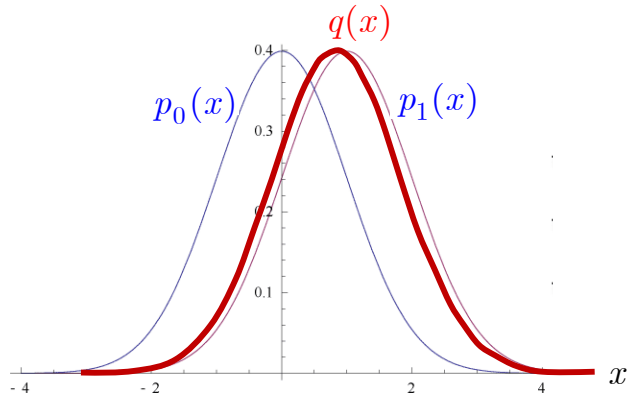
- Conclusion

$$\frac{1}{N}\sum_{i=1}^{N}\log\frac{p_1(x_i)}{p_0(x_i)}\underset{H_0}{\overset{H_1}{\gtrless}}\ln\gamma \quad\Rightarrow\quad D_{\mathrm{KL}}(q\parallel p_0)-D_{\mathrm{KL}}(q\parallel p_1)\underset{H_0}{\overset{H_1}{\gtrless}}\ln\gamma$$

$$D_{\mathrm{KL}}(q\parallel p_0)\underset{H_0}{\overset{H_1}{\gtrless}}D_{\mathrm{KL}}(q\parallel p_1)$$



$D_{\mathrm{KL}}(q\parallel p_0)<D_{\mathrm{KL}}(q\parallel p_1)\Rightarrow H_0$



$D_{\mathrm{KL}}(q\parallel p_0)>D_{\mathrm{KL}}(q\parallel p_1)\Rightarrow H_1$

Continuous R.V.: conditional relative entropy

Definition (conditional relative entropy): *Conditional relative entropy* between pdfs $p(y|x)$ and $q(y|x)$:

$$D_{\mathrm{KL}}\left(p(y|x)\middle\|q(y|x)\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x)p(y|x)\log\left(\frac{p(y|x)}{q(y|x)}\right)dx\,dy$$

$$= \underbrace{E_{p(x)}E_{p(y|x)}}_{E_{p(x,y)}}\left[\log_2\frac{p(y|x)}{q(y|x)}\right]$$

$$= E_{p(x,y)}\left[\log_2\frac{p(y|x)}{q(y|x)}\right]$$

Continuous R.V.: chain rule for relative entropy

- Chain rule

$$D_{\mathrm{KL}}\left(p\left(x,y\right)\middle\|q\left(x,y\right)\right) = D_{\mathrm{KL}}\left(p\left(x\right)\middle\|q\left(x\right)\right) + D_{\mathrm{KL}}\left(p\left(y\middle|x\right)\middle\|q\left(y\middle|x\right)\right)$$

$$= \int\limits_{x\in\mathcal{X}}\int\limits_{y\in\mathcal{Y}} p\left(y,x\right)\log_2\frac{p\left(y\middle|x\right)p(x)}{q\left(y\middle|x\right)q(x)}\,dx\,dy =$$

$$= \int\limits_{x\in\mathcal{X}}\int\limits_{y\in\mathcal{Y}} p\left(y,x\right)\left(\log_2\frac{p\left(y\middle|x\right)}{q\left(y\middle|x\right)} + \log_2\frac{p(x)}{q(x)}\right)dx\,dy =$$

$$= \int\limits_{x\in\mathcal{X}}\int\limits_{y\in\mathcal{Y}} p\left(y,x\right)\log_2\frac{p\left(y\middle|x\right)}{q\left(y\middle|x\right)}\,dx\,dy + \int\limits_{x\in\mathcal{X}}\int\limits_{y\in\mathcal{Y}} p\left(y,x\right)\log_2\frac{p(x)}{q(x)}\,dx\,dy =$$

$$= D_{\mathrm{KL}}\left(p\left(x\right)\middle\|q\left(x\right)\right) + D_{\mathrm{KL}}\left(p\left(y\middle|x\right)\middle\|q\left(y\middle|x\right)\right)$$

Continuous R.Vs.: Mutual information

Definition (muutal information): *Mutual information* for continuous random variable has the same properties as for the discrete random variablea

$$I\left(X;Y\right) = \int_{\mathcal{X}}\int_{\mathcal{Y}} f_{X,Y}\left(x,y\right)\log_2\frac{f_{X,Y}\left(x,y\right)}{f_X\left(x\right)f_Y\left(y\right)}dxdy$$
$$= D_{\mathrm{KL}}\left(f_{X,Y}\left(x,y\right)\middle\|f_X\left(x\right)f_Y\left(y\right)\right)$$

$$I\left(X;Y\right) \geq 0$$
$$I\left(X;Y\right) = h\left(X\right) - h\left(X\middle|Y\right)$$
$$I\left(X;Y\right) = h\left(Y\right) - h\left(Y\middle|X\right)$$
$$I\left(X;Y\right) = h\left(X\right) + h\left(Y\right) - h\left(X,Y\right)$$

Continuous R.Vs.: Mutual information

$$I(X;Y) = D_{\mathrm{KL}}\left(f_{X,Y}(x,y)\middle\|f_X(x)f_Y(y)\right)$$
$$= E_{f_Y(y)}\left[D_{\mathrm{KL}}\left(f_{X|Y=y}(x|y)\middle\|f_X(x)\right)\right] = E_{f_X(x)}\left[D_{\mathrm{KL}}\left(f_{Y|X=x}(y|x)\middle\|f_Y(x)\right)\right]$$

- Proof

$$I(X;Y) = \frac{\sum_{y\in\mathcal{Y}} f_Y(y)\sum_{x\in\mathcal{X}} f_{X|Y=y}(x|y)\log\frac{f_{X|Y=y}(x|y)}{f_X(x)}}{\sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} f_{X,Y}(x,y)}$$
$$= \sum_{y\in\mathcal{Y}} f_Y(y) D_{\mathrm{KL}}\left(f_{X|Y=y}(x|y)\parallel f_X(x)\right)$$
$$= E_{f_Y(y)}\left[D_{\mathrm{KL}}\left(f_{X|Y=y}(x|y)\middle\|f_X(x)\right)\right]$$

Remark: here we highlight the fact of conditioning in the KLD. However, many ML works simplify it to the notations on the previous pages.

Continuous R.Vs.: Mutual information

- Conditional mutual information

$$I\left(X;Y\middle|Z\right)=H\left(X\middle|Z\right)-H\left(X\middle|Y,Z\right)$$

  - Proof

$$I\left(X;Y\middle|Z\right)=E_{f\left(x,y,z\right)}\left[\log\frac{f_{X,Y|Z}\left(x,y\middle|z\right)}{f_{X|Z}\left(x\middle|z\right)f_{Y|Z}\left(y\middle|z\right)}\right]=E_{f\left(x,y,z\right)}\left[\log\frac{f_{X|Y,Z}\left(x\middle|y,z\right)}{f_{X|Z}\left(x\middle|z\right)}\right]$$

$$f_{X,Y|Z}\left(x,y\middle|z\right)=f_{Y|Z}\left(y\middle|z\right)f_{X|Y,Z}\left(x\middle|y,z\right)$$

$$=\underbrace{E_{f\left(x,y,z\right)}\left[\log f_{X|Y,Z}\left(x\middle|y,z\right)\right]}_{-h\left(X\middle|Y,Z\right)}-\underbrace{E_{f\left(x,y,z\right)}\left[\log f_{X|Z}\left(x\middle|z\right)\right]}_{h\left(X\middle|Z\right)}$$

- Chain rule

$$I\left(X_1,X_2,\ldots,X_N;Y\right)=\sum_{i=1}^{N}I(X_i;Y\middle|X_{i-1},\ldots,X_1)$$

Continuous R.Vs.: MI for univariate Gaussians

- Consider a transformation of Gaussian random variable $X \sim \mathcal{N}\left(0, \sigma_X^2\right)$ by adding an independent random variable $Z \sim \mathcal{N}\left(0, \sigma_Z^2\right)$, i.e., $Y = X + Z$

$$X \sim \mathcal{N}\left(0, \sigma_X^2\right) \longrightarrow \boxed{p\left(y \middle| x\right)} \longrightarrow Y \sim \mathcal{N}\left(0, \underbrace{\sigma_X^2 + \sigma_Z^2}_{\sigma_Y^2}\right)$$

$$y = x + z$$

$$Z \sim \mathcal{N}\left(0, \sigma_Z^2\right)$$

- Compute mutual information via entropy and show the equivalence of all forms

$$I\left(X;Y\right) = h\left(Y\right) - h\left(Y\middle|X\right) = \frac{1}{2}\log\left(1 + \frac{\sigma_X^2}{\sigma_Z^2}\right)$$

$$I\left(X;Y\right) = h\left(X\right) - h\left(X\middle|Y\right) = \frac{1}{2}\log\left(1 + \frac{\sigma_X^2}{\sigma_Z^2}\right) \qquad I\left(X;Y\right) = D_{\mathrm{KL}}(p(x,y)\Big\|p(x)p(y))$$

$$I\left(X;Y\right) = h\left(X\right) + h\left(Y\right) - h\left(X,Y\right) = \frac{1}{2}\log\left(1 + \frac{\sigma_X^2}{\sigma_Z^2}\right)$$

Continuous R.Vs.: MI for univariate Gaussians

$$I(X;Y) = h(Y) - h(Y|X) = \frac{1}{2}\log\left(1 + \frac{\sigma_X^2}{\sigma_Z^2}\right)$$

- Proof
  - Random variable $Y \sim \mathcal{N}\left(0, \underbrace{\sigma_X^2 + \sigma_Z^2}_{\sigma_Y^2}\right)$

$$h(Y) = \frac{1}{2}\log\left(2\pi e\sigma_Y^2\right) = \frac{1}{2}\log\left(2\pi e(\sigma_X^2 + \sigma_Z^2)\right)$$

  - Since $Y = X + Z$

$$h(Y|X) = h(X + Z|X) = h(Z) = \frac{1}{2}\log\left(2\pi e\sigma_Z^2\right)$$

  - Finally

$$I(X;Y) = h(Y) - h(Y|X) = \frac{1}{2}\log\left(\frac{2\pi e(\sigma_X^2 + \sigma_Z^2)}{2\pi e\sigma_Z^2}\right) = \frac{1}{2}\log\left(1 + \frac{\sigma_X^2}{\sigma_Z^2}\right)$$

Proof: Case 1

Continuous R.Vs.: MI for univariate Gaussians

$$I(X;Y) = h(X) - h(X|Y) = \frac{1}{2}\log\left(1 + \frac{\sigma_X^2}{\sigma_Z^2}\right)$$

- Proof
  - Random variable $X \sim \mathcal{N}(0, \sigma_X^2)$

$$h(X) = \frac{1}{2}\log\left(2\pi e\sigma_X^2\right) = \frac{1}{2}\log\left(2\pi e\sigma_X^2\right)$$

  - Since $Y = X + Z$

$$h(X|Y) = h(X|X+Z) = \frac{1}{2}\log\left(2\pi e\sigma_{X|Y}^2\right)$$

$$\text{where } \sigma_{X|Y}^2 = \frac{\sigma_X^2\sigma_Z^2}{\sigma_X^2 + \sigma_Z^2}$$

  - Finally

$$I(X;Y) = h(X) - h(X|Y) = \frac{1}{2}\log\left(\frac{2\pi e(\sigma_X^2 + \sigma_Z^2)}{2\pi e\sigma_Z^2}\right) = \frac{1}{2}\log\left(1 + \frac{\sigma_X^2}{\sigma_Z^2}\right)$$

Proof: Case 2

Continuous R.Vs.: MI for univariate Gaussians

- Estimation of X given $Y = X + Z$ based on maximum a posteriori (MAP) principle

$$\hat{x} = \arg\max_x p(y|x)p(x)$$

$$p(y|x) = \mathcal{N}\left(x, \sigma_Z^2\right) = \frac{1}{\sqrt{2\pi\sigma_Z^2}} e^{-\frac{(y-x)^2}{2\sigma_Z^2}} \qquad p(x) = \mathcal{N}\left(0, \sigma_X^2\right) = \frac{1}{\sqrt{2\pi\sigma_X^2}} e^{-\frac{(x-0)^2}{2\sigma_X^2}}$$

$$\hat{x} = \arg\max_x \log\left[p(y|x)p(x)\right] \Rightarrow \frac{d}{dx}\left[\log p(y|x) + \log p(x)\right] = 0$$

$$\frac{d}{dx}\left[-\frac{(y-x)^2}{2\sigma_Z^2} - \frac{(x)^2}{2\sigma_X^2}\right] = 0 \Rightarrow \frac{2}{2\sigma_Z^2}(y-x) - \frac{2x}{2\sigma_X^2} = 0$$

$$\frac{y}{\sigma_Z^2} - \left(\frac{1}{\sigma_Z^2} + \frac{1}{\sigma_X^2}\right)x = 0 \Rightarrow \hat{x} = \frac{\sigma_Z^2\sigma_X^2}{\sigma_Z^2 + \sigma_X^2}\frac{1}{\sigma_Z^2}y = \frac{\sigma_X^2}{\sigma_Z^2 + \sigma_X^2}y$$

Continuous R.Vs.: MI for univariate Gaussians

- Variance of MAP estimation

$$\sigma^2_{X|Y} = E_{p(x)}\left[\left|X - \hat{X}\right|^2\right] = \frac{\sigma_X^2\sigma_Z^2}{\sigma_X^2 + \sigma_Z^2}$$

  - Proof

$$\sigma^2_{X|Y} = E_{p(x,z)}\left[\left|X - \hat{X}\right|^2\right] = E_{p(x,z)}\left[\left|X - \frac{\sigma_X^2}{\sigma_X^2 + \sigma_Z^2}Y\right|^2\right]$$

$$= E_{p(x,z)}\left[\left|X - \frac{\sigma_X^2}{\sigma_X^2 + \sigma_Z^2}(X+Z)\right|^2\right] = E_{p(x,z)}\left[\left|\frac{\sigma_Z^2}{\sigma_X^2 + \sigma_Z^2}X - \frac{\sigma_X^2}{\sigma_X^2 + \sigma_Z^2}Z\right|^2\right]$$

$$= \left(\frac{\sigma_Z^2}{\sigma_X^2 + \sigma_Z^2}\right)^2\sigma_X^2 + \left(\frac{\sigma_X^2}{\sigma_X^2 + \sigma_Z^2}\right)^2\sigma_Z^2 = \frac{\sigma_X^2\sigma_Z^2\left(\sigma_Z^2 + \sigma_X^2\right)}{\left(\sigma_X^2 + \sigma_Z^2\right)^2} = \frac{\sigma_X^2\sigma_Z^2}{\sigma_X^2 + \sigma_Z^2}$$

Continuous R.Vs.: MI for univariate Gaussians

$$I(X;Y) = h(X) + h(Y) - h(X,Y) = \frac{1}{2}\log\left(1 + \frac{\sigma_X^2}{\sigma_Z^2}\right) = \frac{1}{2}\log\left(\frac{1}{1-\rho_{XY}^2}\right)$$

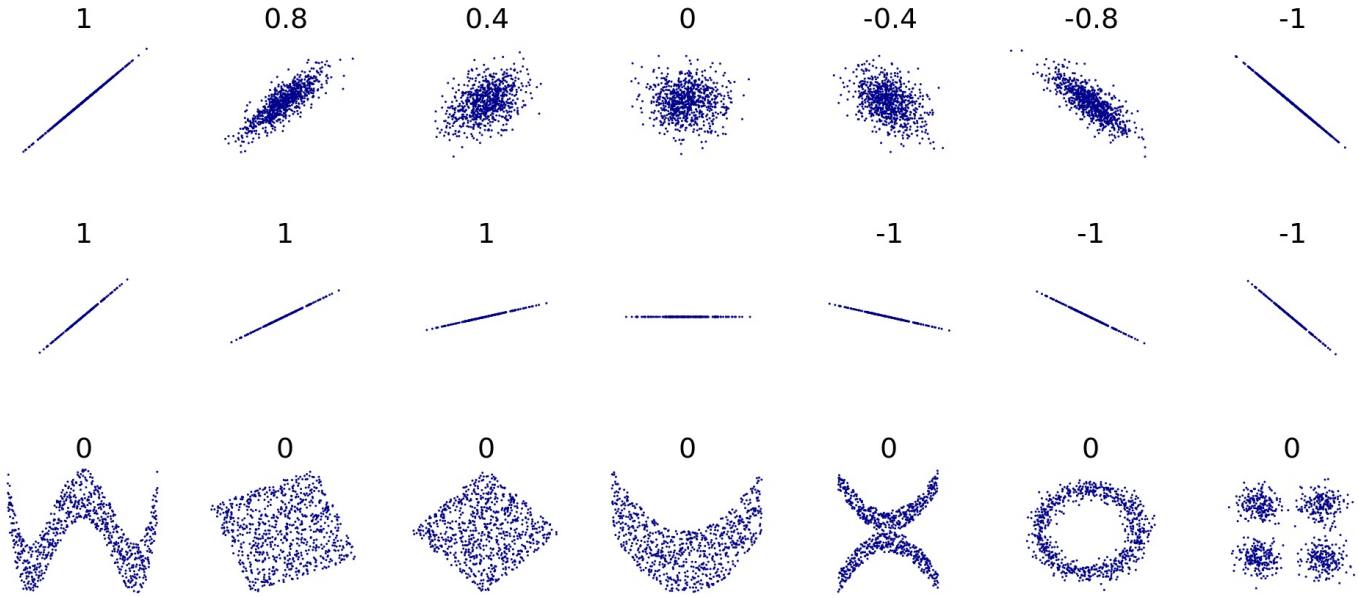
- Remark
  - This exercise is to show not only the equivalence of three forms of mutual information decomposition via the entropy but also to demonstrate:
    - How to compute and use the joint entropy
    - How to link it to the correlation coefficient
  - Note: mutual information vs correlation coefficient (or MSE)
    - Mutual information contains information about **all** dependencies – linear and nonlinear
    - Correlation coefficient measures **only** linear dependencies
  - Note: in the bivariate Gaussian case between X and Y they coincide

Proof: Case 3

https://twitter.com/docmilanfar/status/1400276401513648143?s=20

Continuous R.Vs.: MI for univariate Gaussians

- Correlation coefficient for different data manifolds (pdfs)



$$\rho_{XY} = \frac{Cov[X,Y]}{\sqrt{Var[X]}\sqrt{Var[Y]}}$$

- Note that the correlation reflects the strength and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom).
- N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of Y is zero.

https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Continuous R.Vs.: MI for univariate Gaussians

$$I(X;Y) = h(X) + h(Y) - h(X,Y) = \frac{1}{2}\log\left(1 + \frac{\sigma_X^2}{\sigma_Z^2}\right) = \frac{1}{2}\log\left(\frac{1}{1-\rho_{XY}^2}\right)$$

- Proof
  - Random variable $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$

$$h(X) = \frac{1}{2}\log\left(2\pi e\sigma_X^2\right)$$

$$h(Y) = \frac{1}{2}\log\left(2\pi e\sigma_Y^2\right)$$

  - To find $h(X,Y)$, we will define a joint pdf:

$(X,Y) \sim p(x,y) = \mathcal{N}(\underline{\mu}, \mathbf{K}_{\mathbf{xy}})$ which is Gaussian since $Y = X + Z$

    - where

$$\underline{\mu} = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix} \qquad \mathbf{K}_{\mathbf{xy}} = \begin{bmatrix} \sigma_X^2 & \rho_{XY}\sigma_X\sigma_Y \\ \rho_{XY}\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}$$

Proof: Case 3

Continuous R.Vs.: MI for univariate Gaussians

- With the correlation coefficient

$$\rho_{XY} = \frac{Cov\left[X,Y\right]}{\sqrt{Var\left[X\right]}\sqrt{Var\left[Y\right]}} = \frac{E\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sigma_X\sigma_Y} = \frac{E\left[XY\right]-\mu_X\mu_Y}{\sigma_X\sigma_Y}$$

$$= \frac{E\left[X(X+Z)\right]-\mu_X(\mu_X+0)}{\sigma_X\sigma_Y} = \frac{E\left[X^2\right]-\mu_X^2}{\sigma_X\sigma_Y}$$

$$= \frac{\sigma_X^2}{\sigma_X\sigma_Y} = \frac{\sigma_X}{\sigma_Y} = \pm\sqrt{\frac{\sigma_X^2}{\sigma_X^2+\sigma_Z^2}}$$

- Thus $h\left(X,Y\right) = -E_{p(x,y)}\left[\log_2 p\left(x,y\right)\right] = \frac{1}{2}\log\left((2\pi e)^2 \det\left|\mathbf{K}_{\mathbf{xy}}\right|\right)$

$$\det\left|\mathbf{K}_{\mathbf{xy}}\right| = \det\left\|\begin{bmatrix}\sigma_X^2 & \rho_{XY}\sigma_X\sigma_Y \\ \rho_{XY}\sigma_X\sigma_Y & \sigma_Y^2\end{bmatrix}\right\|$$

$$= \sigma_X^2\sigma_Y^2 - \rho_{XY}^2\sigma_X^2\sigma_Y^2 = (1-\rho_{XY}^2)\sigma_X^2\sigma_Y^2 = \sigma_X^2\sigma_Z^2$$

Proof: Case 3

Continuous R.Vs.: MI for univariate Gaussians

- Substituting

$$I(X;Y) = h(X) + h(Y) - h(X,Y)$$

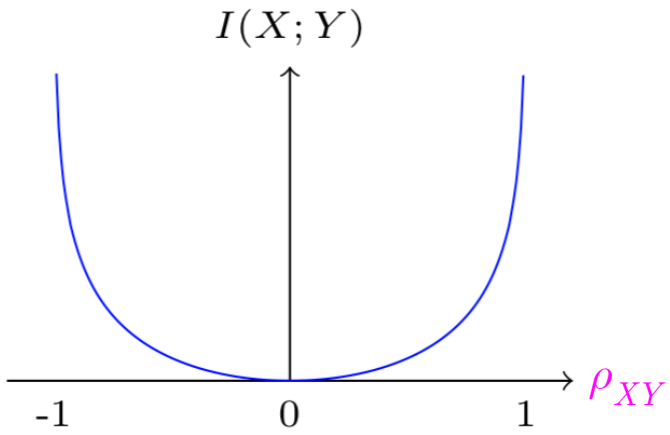
$$= \frac{1}{2}\log\left(2\pi e\sigma_X^2\right) + \frac{1}{2}\log\left(2\pi e \underbrace{\sigma_Y^2}_{\sigma_X^2+\sigma_Z^2}\right) - \frac{1}{2}\log\left((2\pi e)^2\sigma_X^2\sigma_Z^2\right)$$

$$= \frac{1}{2}\log\left(\frac{(2\pi e)^2}{(2\pi e)^2}\frac{\sigma_X^2(\sigma_X^2+\sigma_Z^2)}{\sigma_X^2\sigma_Z^2}\right) = \frac{1}{2}\log\left(1+\frac{\sigma_X^2}{\sigma_Z^2}\right) = \frac{1}{2}\log\left(\frac{1}{1-\rho_{XY}^2}\right)$$

$$\Rightarrow \rho_{XY}^2 = 1 - e^{-2I(X;Y)}$$

$$I(X;Y) = 0 : \rho_{XY}^2 = 1 - e^0 = 0$$

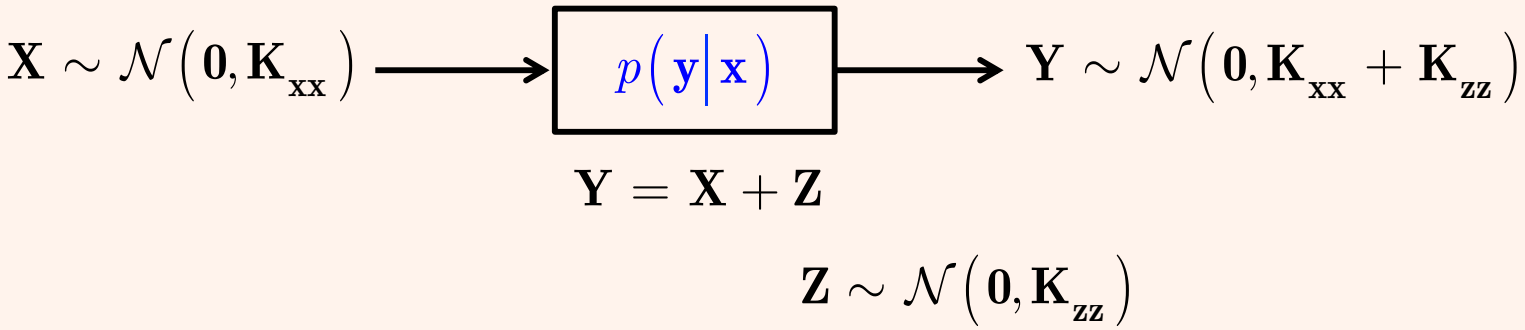
$$I(X;Y) \to \infty : \rho_{XY}^2 \to 1$$



Proof: Case 3

Continuous R.Vs.: MI for multivariate Gaussians

- Consider a transformation of Gaussian random variable $\mathbf{X} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{K}_{\mathbf{xx}}\right)$ by adding an independent random variable $\mathbf{Z} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{K}_{\mathbf{zz}}\right)$, i.e., $\mathbf{Y} = \mathbf{X} + \mathbf{Z}$



- Compute mutual information via entropy and show the symmetry

$$I\left(\mathbf{X};\mathbf{Y}\right) = h\left(\mathbf{Y}\right) - h\left(\mathbf{Y}\middle|\mathbf{X}\right) = \frac{1}{2}\log\left(\frac{\det\left|\mathbf{K}_{\mathbf{xx}} + \mathbf{K}_{\mathbf{zz}}\right|}{\det\left|\mathbf{K}_{\mathbf{zz}}\right|}\right)$$

$$I\left(\mathbf{X};\mathbf{Y}\right) = h\left(\mathbf{X}\right) - h\left(\mathbf{X}\middle|\mathbf{Y}\right) = \frac{1}{2}\log\left(\frac{\det\left|\mathbf{K}_{\mathbf{xx}} + \mathbf{K}_{\mathbf{zz}}\right|}{\det\left|\mathbf{K}_{\mathbf{zz}}\right|}\right)$$

Continuous R.Vs.: MI for multivariate Gaussians

$$I\left(\mathbf{X};\mathbf{Y}\right)=h\left(\mathbf{Y}\right)-h\left(\mathbf{Y}\middle|\mathbf{X}\right)=\frac{1}{2}\log\left(\frac{\det\left|\mathbf{K}_{\mathbf{xx}}+\mathbf{K}_{\mathbf{zz}}\right|}{\det\left|\mathbf{K}_{\mathbf{zz}}\right|}\right)$$

- Proof
  - Random variable $\mathbf{Y}\sim\mathcal{N}\left(\mathbf{0},\mathbf{K}_{\mathbf{xx}}+\mathbf{K}_{\mathbf{zz}}\right)$

$$h\left(\mathbf{Y}\right)=\frac{1}{2}\ln\left((2\pi e)^{N}\det\left|\mathbf{K}_{\mathbf{xx}}+\mathbf{K}_{\mathbf{zz}}\right|\right)$$

  - Since $\mathbf{Y}=\mathbf{X}+\mathbf{Z}$

$$h\left(\mathbf{Y}\middle|\mathbf{X}\right)=h\left(\mathbf{X}+\mathbf{Z}\middle|\mathbf{X}\right)=h\left(\mathbf{Z}\right)=\frac{1}{2}\log\left((2\pi e)^{N}\det\left|\mathbf{K}_{\mathbf{zz}}\right|\right)$$

  - Finally

$$I\left(\mathbf{X};\mathbf{Y}\right)=h\left(\mathbf{Y}\right)-h\left(\mathbf{Y}\middle|\mathbf{X}\right)=\frac{1}{2}\log\left(\frac{(2\pi e)^{N}\det\left|\mathbf{K}_{\mathbf{xx}}+\mathbf{K}_{\mathbf{zz}}\right|}{(2\pi e)^{N}\det\left|\mathbf{K}_{\mathbf{zz}}\right|}\right)$$

Continuous R.Vs.: MI for multivariate Gaussians

- Practical simplification

$$I\left(\mathbf{X};\mathbf{Y}\right) = h\left(\mathbf{Y}\right) - h\left(\mathbf{Y}\middle|\mathbf{X}\right) = \frac{1}{2}\log\left(\frac{(2\pi e)^N \det\left|\mathbf{K}_{\mathbf{xx}} + \mathbf{K}_{\mathbf{zz}}\right|}{(2\pi e)^N \det\left|\mathbf{K}_{\mathbf{zz}}\right|}\right)$$

- The term $\det\left|\mathbf{K}_{\mathbf{zz}}\right|$

$$\mathbf{K}_{\mathbf{zz}} = \mathbf{U}\Sigma_{\mathbf{zz}}\mathbf{U}^T, \text{ with } \Sigma_{\mathbf{zz}} = diag\left(\sigma_1,\cdots,\sigma_N\right) \text{ or } \Sigma_{ii} = \sigma_i, \forall i \text{ and } \mathbf{U}\mathbf{U}^T = \mathbf{I}$$

$$\det\left|\mathbf{K}_{\mathbf{zz}}\right| = \left|\det \mathbf{U}\Sigma_{\mathbf{zz}}\mathbf{U}^T\right| = \left|\underbrace{\det \mathbf{U}\mathbf{U}^T}_{\mathbf{I}}\right|\left|\det \Sigma_{\mathbf{zz}}\right| = \left|\det \Sigma_{\mathbf{zz}}\right| = \prod_{i=1}^{N}\sigma_i$$

- The term $\det\left|\mathbf{K}_{\mathbf{xx}} + \mathbf{K}_{\mathbf{zz}}\right|$

$$\det\left|\mathbf{K}_{\mathbf{xx}} + \mathbf{K}_{\mathbf{zz}}\right| = \det\left|\mathbf{K}_{\mathbf{xx}} + \mathbf{U}\Sigma_{\mathbf{zz}}\mathbf{U}^T\right| = \det\left|\mathbf{U}\right| v \det\left|\mathbf{U}^T\right|$$

$$= \det\left|\underbrace{\mathbf{U}^T\mathbf{K}_{\mathbf{xx}}\mathbf{U}}_{\mathbf{A}} + \Sigma_{\mathbf{zz}}\right| = \det\left|\mathbf{A} + \Sigma_{\mathbf{zz}}\right| \le \prod_{i=1}^{N}\left(\mathbf{A}_{ii} + \sigma_i\right)$$

Continuous R.Vs.: MI for multivariate Gaussians

- Practical simplification

$$
\begin{aligned}
I\left(\mathbf{X};\mathbf{Y}\right) = h\left(\mathbf{Y}\right) - h\left(\mathbf{Y}\middle|\mathbf{X}\right) &= \frac{1}{2}\log\left(\frac{(2\pi e)^N \det\left|\mathbf{K}_{\mathbf{xx}} + \mathbf{K}_{\mathbf{zz}}\right|}{(2\pi e)^N \det\left|\mathbf{K}_{\mathbf{zz}}\right|}\right) \\
&\leq \frac{1}{2}\log\left(\frac{\prod_{i=1}^{N}\left(\mathbf{A}_{ii} + \sigma_i\right)}{\prod_{i=1}^{N}\sigma_i}\right), \text{ where } \mathbf{A} = \mathbf{U}^T\mathbf{K}_{\mathbf{xx}}\mathbf{U} \\
&= \frac{1}{2}\log\left(\prod_{i=1}^{N}\frac{\left(\mathbf{A}_{ii} + \sigma_i\right)}{\sigma_i}\right) \\
&= \frac{1}{2}\log\left(\prod_{i=1}^{N}\left(1 + \frac{\mathbf{A}_{ii}}{\sigma_i}\right)\right) \\
&= \frac{1}{2}\sum_{i=1}^{N}\log\left(1 + \frac{\mathbf{A}_{ii}}{\sigma_i}\right)
\end{aligned}
$$

Scope

- Information theoretic measures for discrete/continuous variables
  - Entropy
  - Conditional entropy
  - Joint entropy
  - Relative entropy (KL-divergence)
  - Cross entropy
  - Mutual information
  - Additional topics:
    - f-divergence
    - practical computations from samples
      - Estimation of KL- and f-divergences by density ratio trick
      - Variational approximation of mutual information

f-divergences

Definition (f-divergence): f-divergences or Ali-Silvey distances between $p(x)$ and $q(x)$

$$D_f(p \parallel q) = \int_{\mathcal{X}} p(x) f\left(\frac{q(x)}{p(x)}\right) dx$$

where the function (a.k.a. *generator function*) $f : \mathbb{R}_+ \to \mathbb{R}$ is convex and satisfy $f(1) = 0$

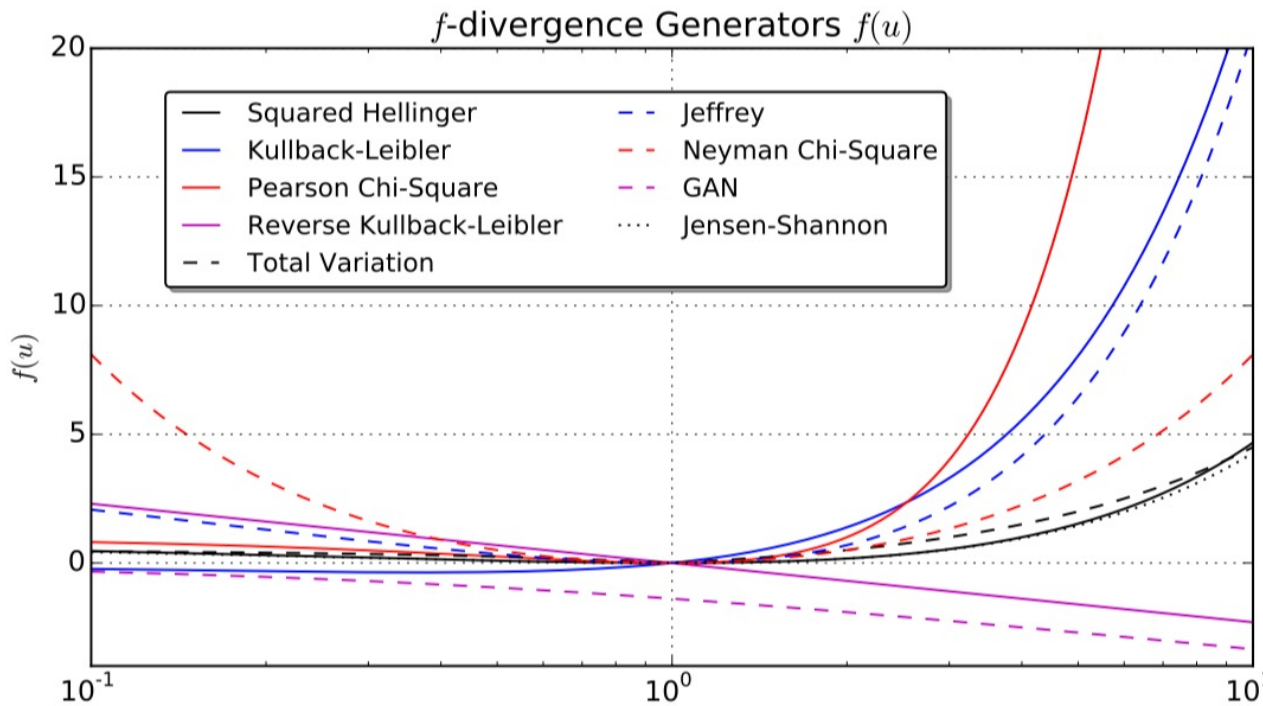
- Recall: KLD

$$D_{\mathrm{KL}}(p \parallel q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx = E_{p(x)}\left[\log_2 \frac{p(x)}{q(x)}\right]$$

- Remark: many works define $D_f(p \parallel q) = \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) dx$

f-divergences

- Different choices of $f(u)$



https://arxiv.org/pdf/1606.00709.pdf

f-divergences

- Different choices of $f(u)$

Name of the f-divergence	Formula $D_f(p \parallel q)$	Generator $f(u)$ with $f(1)=0$
Total variation (metric)	$\frac{1}{2}\int \lvert p(x)-q(x)\rvert \mathrm{d}\nu(x)$	$\frac{1}{2}\lvert u-1\rvert$
Squared Hellinger	$\int (\sqrt{p(x)}-\sqrt{q(x)})^2 \mathrm{d}\nu(x)$	$(\sqrt{u}-1)^2$
Pearson χ^2_P	$\int \frac{(q(x)-p(x))^2}{p(x)} \mathrm{d}\nu(x)$	$(u-1)^2$
Neyman χ^2_N	$\int \frac{(p(x)-q(x))^2}{q(x)} \mathrm{d}\nu(x)$	$\frac{(1-u)^2}{u}$
Pearson-Vajda χ^k_P	$\int \frac{(q(x)-\lambda p(x))^k}{p^{k-1}(x)} \mathrm{d}\nu(x)$	$(u-1)^k$
Pearson-Vajda $\lvert\chi\rvert^k_P$	$\int \frac{\lvert q(x)-\lambda p(x)\rvert^k}{p^{k-1}(x)} \mathrm{d}\nu(x)$	$\lvert u-1\rvert^k$
Kullback-Leibler	$\int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}\nu(x)$	$-\log u$
reverse Kullback-Leibler	$\int q(x) \log \frac{q(x)}{p(x)} \mathrm{d}\nu(x)$	$u \log u$
α-divergence	$\frac{4}{1-\alpha^2}(1-\int p^{\frac{1-\alpha}{2}}(x) q^{1+\alpha}(x) \mathrm{d}\nu(x))$	$\frac{4}{1-\alpha^2}(1-u^{\frac{1+\alpha}{2}})$
Jensen-Shannon	$\frac{1}{2}\int (p(x) \log \frac{2p(x)}{p(x)+q(x)} + q(x) \log \frac{2q(x)}{p(x)+q(x)}) \mathrm{d}\nu(x)$	$-(u+1)\log \frac{1+u}{2} + u \log u$

https://arxiv.org/pdf/1309.3029.pdf

f-divergences: TV, Pearson and KLD

- **Examples** $D_f(p \parallel q) = \int_{\mathcal{X}} p(x) f\left(\frac{q(x)}{p(x)}\right) dx$
  - Total variation distance $f(u) = \frac{1}{2}|u-1|$

$$D_{\mathrm{TV}}(p \parallel q) = \frac{1}{2}\int_{\mathcal{X}} p(x)\left|\frac{q(x)}{p(x)} - 1\right| dx = \frac{1}{2}\int_{\mathcal{X}} \left|q(x) - p(x)\right| dx$$

  - Pearson χ^2 – divergence $f(u) = (u-1)^2$

$$D_{\chi^2}(p \parallel q) = \int_{\mathcal{X}} p(x)\left(\frac{q(x)}{p(x)} - 1\right)^2 dx = \int_{\mathcal{X}} \frac{\left(q(x) - p(x)\right)^2}{p(x)} dx$$

  - Kullback-Leibler divergence $f(u) = -\log u$

$$D_{\mathrm{KL}}(p \parallel q) = -\int_{\mathcal{X}} p(x) \log \frac{q(x)}{p(x)} dx = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} dx$$

f-divergences: useful inequality

- **TV and KLD**

$$D_{\mathrm{TV}}(p \parallel q)^2 = \left\| p(x) - q(x) \right\|_1^2 \leq \frac{1}{2 \log e} D_{\mathrm{KL}}(p \parallel q)$$

- Additionally

$$D_{\mathrm{TV}}(p \parallel q) = \frac{1}{2} \int_{\mathcal{X}} p(x) \left| \frac{q(x)}{p(x)} - 1 \right| dx = \frac{1}{2} \int_{\mathcal{X}} \left| q(x) - p(x) \right| dx$$

and

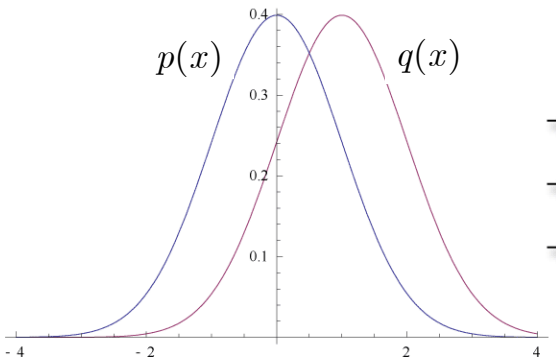
$$\frac{1}{2} \mid a - b \mid = \frac{a+b}{2} - \min(a,b) = \max(a,b) - \frac{a+b}{2}$$

$$\begin{aligned} D_{\mathrm{TV}}(p \parallel q) &= \int_{\mathcal{X}} \left(\frac{p(x) + q(x)}{2} - \min(p(x), q(x)) \right) dx \\ &= 1 - \int_{\mathcal{X}} \min(p(x), q(x)) dx = \int_{\mathcal{X}} \max(p(x), q(x)) dx - 1 \end{aligned}$$

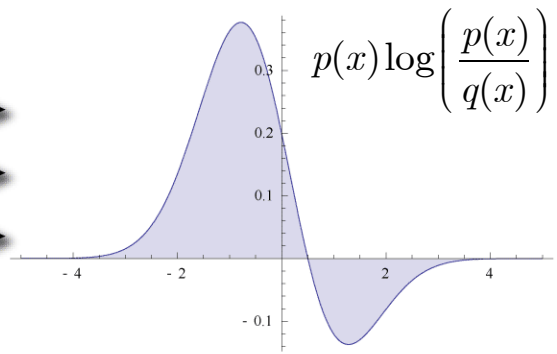
f-divergences: useful inequality

- TV and KLD

$$D_{\mathrm{KL}}(p \parallel q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} dx$$

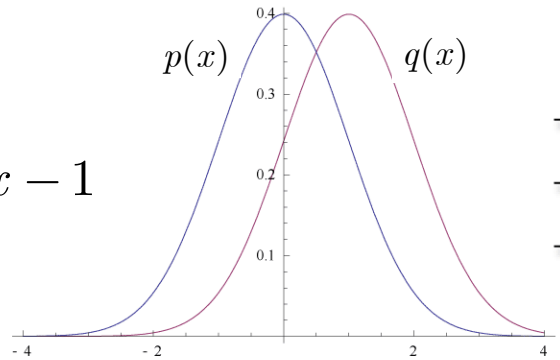


Original Gaussian PDF's

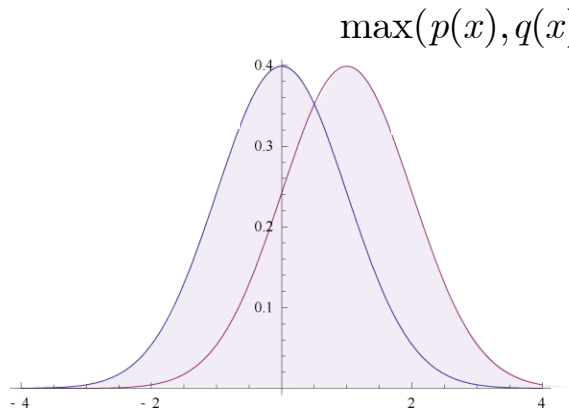


KL Area to be Integrated $D_{\mathrm{KL}}(p \parallel q)$

$$D_{\mathrm{TV}}(p \parallel q) = \int_{\mathcal{X}} \max(p(x), q(x)) dx - 1$$



Original Gaussian PDF's



Area to be Integrated $D_{\mathrm{TV}}(p \parallel q)$

f-divergences: useful inequality

- Also a useful inequality

$$D_{\mathrm{TV}}(p \parallel q) = \int_{\mathcal{X}} \left(\frac{p(x)+q(x)}{2} - \min(p(x), q(x)) \right) dx$$
$$= 1 - \int_{\mathcal{X}} \min(p(x), q(x)) dx = \int_{\mathcal{X}} \max(p(x), q(x)) dx - 1$$

$$\min(a, b) \leq a^{\alpha} b^{1-\alpha}, \forall a, b > 0$$

$$\min(p(x), q(x)) \leq p(x)^{\alpha} q(x)^{1-\alpha}$$

$$D_{\mathrm{TV}}(p \parallel q) = 1 - \int_{\mathcal{X}} \min(p(x), q(x)) dx \geq 1 - \int_{\mathcal{X}} p(x)^{\alpha} q(x)^{1-\alpha} dx$$

f-divergences: Jensen-Shannon

- **Examples**
  - Jensen-Shannon divergence $f(u) = -(u+1)\log\frac{1+u}{2} + u\log u$

$$D_{\mathrm{JS}}(p \parallel q) = \frac{1}{2}\int\left(p(x)\log\frac{2p(x)}{p(x)+q(x)} + q(x)\log\frac{2q(x)}{p(x)+q(x)}\right)dx$$

  - Alternative definition: symmetric KLD

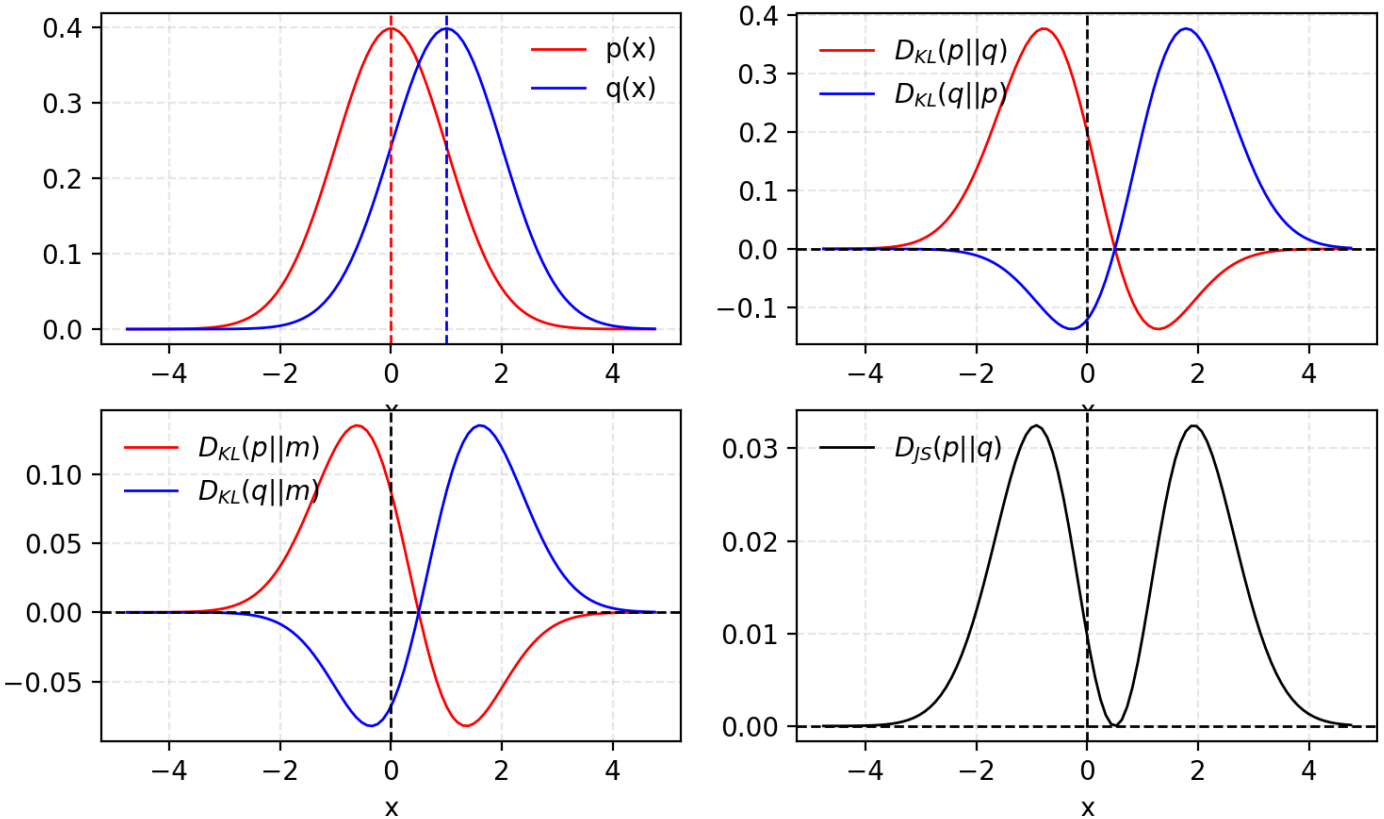
$$D_{\mathrm{JS}}(p \parallel q) = \frac{1}{2}D_{\mathrm{KL}}(p \parallel m) + \frac{1}{2}D_{\mathrm{KL}}(q \parallel m) \qquad m(x) = \frac{1}{2}\big(p(x)+q(x)\big)$$

$$D_{\mathrm{KL}}(p \parallel m) = \int_{\mathcal{X}} p(x)\log\frac{2p(x)}{p(x)+q(x)}dx$$

$$D_{\mathrm{KL}}(q \parallel m) = \int_{\mathcal{X}} q(x)\log\frac{2q(x)}{p(x)+q(x)}dx$$

f-divergences: Jensen-Shannon

$$p(x) = \mathcal{N}(0,1) \qquad q(x) = \mathcal{N}(1,1) \qquad m(x) = \frac{1}{2}\big(p(x) + q(x)\big)$$



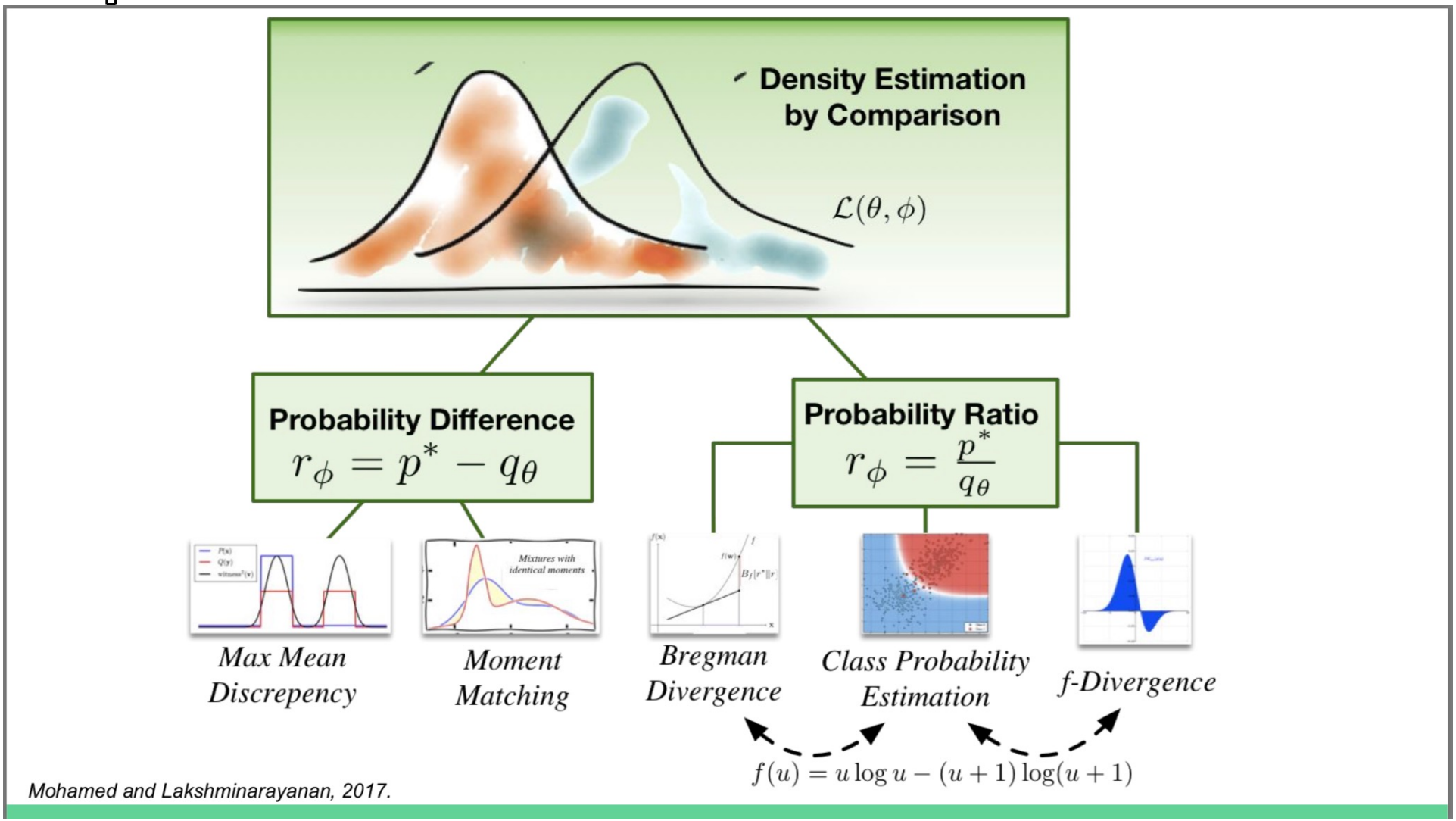
- KL divergence is asymmetric while JS divergence is symmetric

https://lilianweng.github.io/lil-log/2017/08/20/from-GAN-to-WGAN.html

Scope

- Information theoretic measures for discrete/continuous variables
  - Entropy
  - Conditional entropy
  - Joint entropy
  - Relative entropy (KL-divergence)
  - Cross entropy
  - Mutual information
    - Additional topics:
      - f-divergence
      - practical computations from samples
        - Estimation of KL- and f-divergences by density ratio trick
        - Variational approximation of mutual information

f-divergence



https://www.shakirm.com/slides/DeepGenModelsTutorial.pdf

Estimation of density ratio for KL divergence

- KL divergence between distributions $p(x)$ and $q(x)$ is defined as

$$D_{\mathrm{KL}}(p \parallel q) = \int_{\mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$$

- It can be expressed in terms of density ratio $r(x) = \dfrac{p(x)}{q(x)}$

$$D_{\mathrm{KL}}(p \parallel q) = \int_{\mathcal{X}} p(x) \log r(x) dx = E_{p(x)}\left[\log r(x)\right]$$

  - Remark: similar situation for f-divergence

$$D_f(p \parallel q) = \int_{\mathcal{X}} p(x) f\left(\frac{q(x)}{p(x)}\right) dx = \int_{\mathcal{X}} p(x) f\left(r^{-1}(x)\right) dx = E_{p(x)}\left[f\left(r^{-1}(x)\right)\right]$$

Estimation of density ratio for KL divergence

- **Challenges**
  - In both KL- or f-divergence cases

$$D_{\mathrm{KL}}(p \parallel q) = E_{p(x)}\left[\log r(x)\right] \quad \text{and} \quad D_f(p \parallel q) = E_{p(x)}\left[f\left(r^{-1}(x)\right)\right]$$

we need to know how to compute

- Expectation $E_{p(x)}\left[.\right]$ wrt $p(x)$
- Density ratio $r(x) = \dfrac{p(x)}{q(x)}$

Estimation of density ratio for KL divergence

- Practical issues:

Case 1

- Suppose $p(x)$ and $q(x)$ are known (explicitly) but
  - It is difficult to compute mathematically the expectation $E_{p(x)}[.]$ but still possible to compute the ratio $r(x) = \dfrac{p(x)}{q(x)}$
  - It is intractable to compute the ratio $r(x) = \dfrac{p(x)}{q(x)}$

Case 2

- Suppose $p(x)$ and $q(x)$ are not know explicitly but we know some samples generated from these distribution (univariate or multivariate)

$$x_p^{(1)}, \ldots, x_p^{(N_p)} \sim p(x) \qquad x_q^{(1)}, \ldots, x_q^{(N_q)} \sim q(x)$$

implicit distributions

- How to estimate $D_{\mathrm{KL}}(p \parallel q)$ or $D_f(p \parallel q)$ in these cases?

Estimation of density ratio for KL divergence

- Case 1:
  - Monte Carlo estimation

$$E_{p(x)}\left[.\right] \rightarrow \frac{1}{N}\sum_{i=1}^{M}$$

  - KL-divergence with analytically tractable density ratio $r(x) = \frac{p(x)}{q(x)}$

$$D_{\mathrm{KL}}(p \parallel q) = E_{p(x)}\left[\log r(x)\right]$$

$$\approx \frac{1}{M}\sum_{i=1}^{M} \log r(x_p^{(i)}),\ x_p^{(i)} \sim p(x)$$

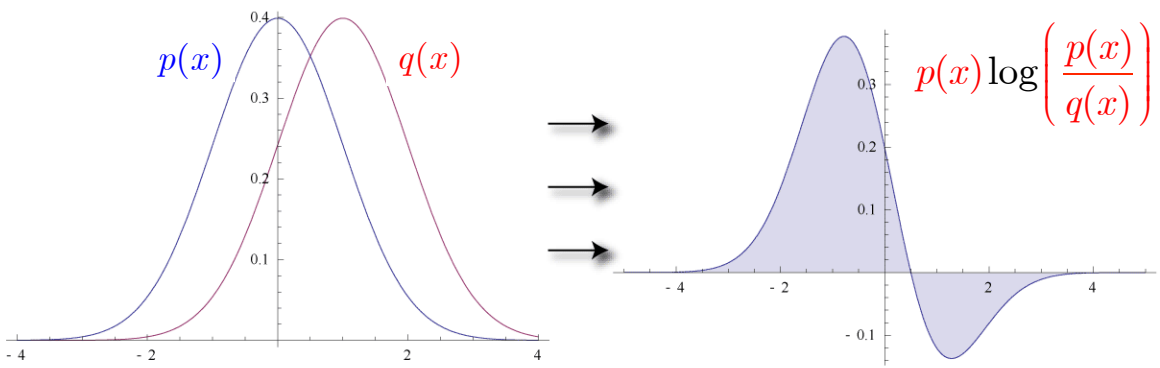
  - f-divergence with analytically tractable density ratio $r(x) = \frac{p(x)}{q(x)}$

$$D_f(p \parallel q) = E_{p(x)}\left[\log\left(r(x)\right)^{-1}\right]$$

$$\approx \frac{1}{M}\sum_{i=1}^{M} \log\left(r(x_p^{(i)})\right)^{-1},\ x_p^{(i)} \sim p(x)$$

Estimation of density ratio for KL divergence

- Case 2:
  - Density ratio estimation – implicit distributions
    - It is intractable to compute the ratio $r(x) = \frac{p(x)}{q(x)}$
    - Distributions are unknown
  - Note: we do not need to know distributions explicitly but we only need to know:
    - either their ratio $r(x) = \frac{p(x)}{q(x)}$ like in KL or JS cases
    - or difference $p(x) - q(x)$ like in Pearson or TV cases



Estimation of density ratio for KL divergence

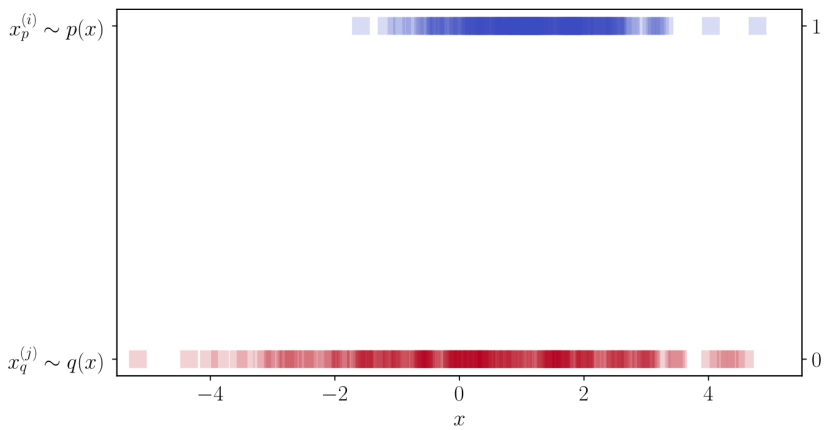
- Case 2:
  - Suppose we have samples from both distributions

$$x_p^{(1)},\ldots,x_p^{(N_p)} \sim p(x) \qquad x_q^{(1)},\ldots,x_q^{(N_q)} \sim q(x)$$

  - Then, we form the dataset $\left\{x_n, y_n\right\}_{n=1}^{N}$, where $N = N_p + N_q$
  - We label samples drawn from $p(x)$ as 1 and those drawn from $q(x)$ as 0

$$\left(x_1,\ldots,x_N\right) = \left(x_p^{(1)},\ldots,x_p^{(N_p)}, x_q^{(1)},\ldots,x_q^{(N_q)}\right)$$

$$\left(y_1,\ldots,y_N\right) = (\underbrace{1,\ldots,1}_{N_p},\underbrace{0,\ldots,0}_{N_q})$$



https://tiao.io/post/density-ratio-estimation-for-kl-divergence-minimization-between-implicit-distributions/

Estimation of density ratio for KL divergence

- Case 2:
  - By the labeling

$$p(x) = \mathcal{P}(x \mid y = 1) \qquad q(x) = \mathcal{P}(x \mid y = 0)$$

  - Using Baye's rule, we can write

$$\mathcal{P}(x \mid y) = \frac{\mathcal{P}(y \mid x)\mathcal{P}(x)}{\mathcal{P}(y)}$$

  - The density ratio can be re-expressed as:

$$r(x) = \frac{p(x)}{q(x)} = \frac{\mathcal{P}(x \mid y = 1)}{\mathcal{P}(x \mid y = 0)}$$

$$= \left(\frac{\mathcal{P}(y = 1 \mid x)\mathcal{P}(x)}{\mathcal{P}(y = 1)} \right) \left(\frac{\mathcal{P}(y = 0 \mid x)\mathcal{P}(x)}{\mathcal{P}(y = 0)} \right)^{-1} = \frac{\mathcal{P}(y = 0)}{\mathcal{P}(y = 1)} \frac{\mathcal{P}(y = 1 \mid x)}{\mathcal{P}(y = 0 \mid x)}$$

Estimation of density ratio for KL divergence

- Case 2:
  - Approximate the ratio of marginal densities by the ratio of sample sizes

$$\frac{\mathcal{P}(y=0)}{\mathcal{P}(y=1)} \approx \frac{N_q}{N_p+N_q}\left(\frac{N_p}{N_p+N_q}\right)^{-1} = \frac{N_q}{N_p}$$

  - Let us assume from now on that $N_q = N_p$. Thus

$$r(x) = \frac{p(x)}{q(x)} = \frac{\mathcal{P}(y=0)}{\mathcal{P}(y=1)}\frac{\mathcal{P}(y=1 \mid x)}{\mathcal{P}(y=0 \mid x)} = \frac{\mathcal{P}(y=1 \mid x)}{\mathcal{P}(y=0 \mid x)}$$

  - This can be developed as

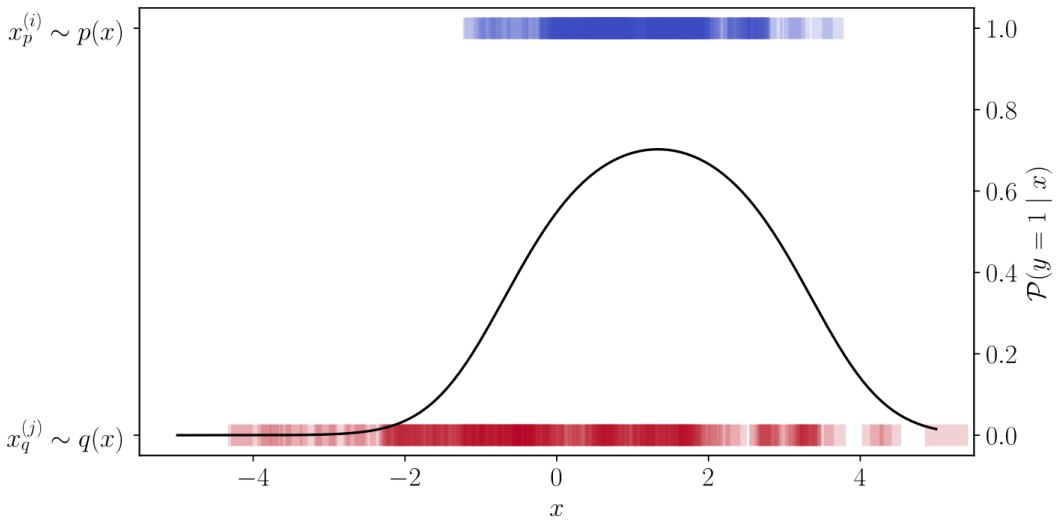
$$r(x) = \frac{\mathcal{P}(y=1 \mid x)}{\mathcal{P}(y=0 \mid x)} = \frac{\mathcal{P}(y=1 \mid x)}{1-\mathcal{P}(y=1 \mid x)}$$
$$= \exp\left[\log\frac{\mathcal{P}(y=1 \mid x)}{1-\mathcal{P}(y=1 \mid x)}\right] = \exp\left[\sigma^{-1}(\mathcal{P}(y=1 \mid x))\right]$$

where $\sigma^{-1}(t) = \log\left(t/(q-t)\right)$ is the *logit* function, or inverse sigmoid function

Estimation of density ratio for KL divergence

- Case 2:
- **Recovering the Class Probability from the Density Ratio**

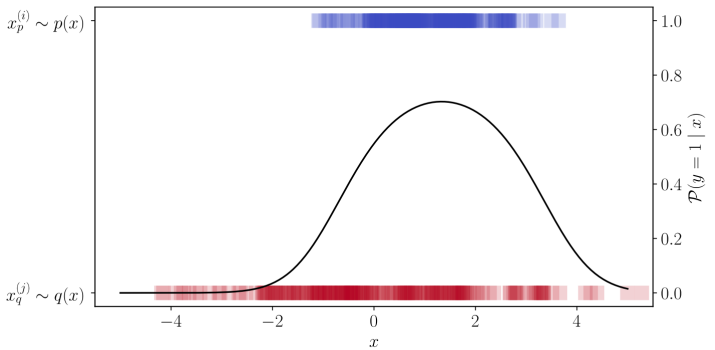
$$\mathcal{P}(y=1 \mid x) = \sigma\Big(\log r(x)\Big) = \frac{p(x)}{p(x)+q(x)}$$

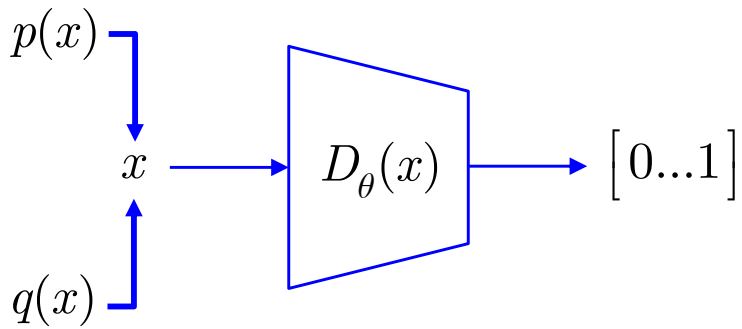


Note: $\mathcal{P}(y=1 \mid x)$ is influenced by both $p(x)$ and $q(x)$ that is a source of potential ambiguity

Estimation of density ratio for KL divergence

- Case 2:
- **Probabilistic Classification with Logistic Regression**
- We can approximate $\mathcal{P}(y=1 \mid x)$ by a parametric function $D_\theta(x)$ with parameters θ





- The corresponding density ratio estimator

$$r_\theta(x) = \exp\Big[\sigma^{-1}\big(D_\theta(x)\big)\Big] \approx \exp\Big[\sigma^{-1}(\mathcal{P}(y=1 \mid x))\Big] = r(x)$$

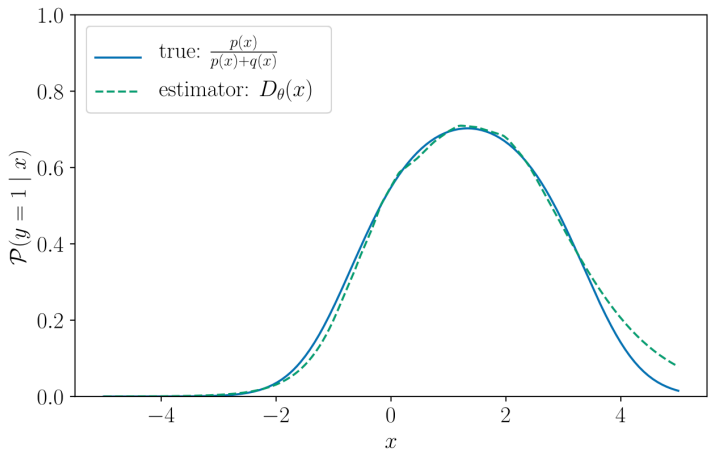
- or alternatively

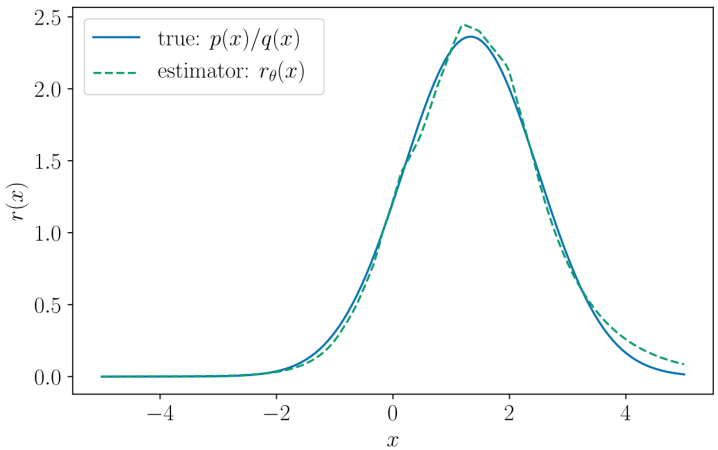
$$D_\theta(x) = \sigma\big(\log r_\theta(x)\big) \approx \sigma\big(\log r(x)\big) = \mathcal{P}(y=1 \mid x)$$

Estimation of density ratio for KL divergence

- Case 2:
- **Training the parameters of estimator** $D_\theta(x)$
  - Using binary cross-entropy, one can try to match the score $D_\theta(x)$ with $p(x)$ and the score $\big(1 - D_\theta(x)\big)$ with $q(x)$

$$\begin{aligned}\mathcal{L}(\theta) &:= -\mathbb{E}_{p(x)}\big[\log D_\theta(x)\big] - \mathbb{E}_{q(x)}\big[\log\big(1 - D_\theta(x)\big)\big] \\ &= -\mathbb{E}_{p(x)}\big[\log\sigma\big(\log r_\theta(x)\big)\big] - \mathbb{E}_{q(x)}\big[\log\big(1 - \sigma\big(\log r_\theta(x)\big)\big)\big]\end{aligned}$$





Estimation of density ratio for KL divergence

- Case 2:
- **Back to Monte Carlo estimation**
  - Having an obtained an estimate of the log density ratio, it is now feasible to perform Monte Carlo estimation

$$D_{\mathrm{KL}}(p \parallel q) = E_{p(x)}\Big[\log r(x)\Big]$$
$$\approx \frac{1}{M}\sum_{i=1}^{M} \log r(x_p^{(i)}),\ x_p^{(i)} \sim p(x)$$
$$\approx \frac{1}{M}\sum_{i=1}^{M} \log r_\theta(x_p^{(i)}),\ x_p^{(i)} \sim p(x)$$

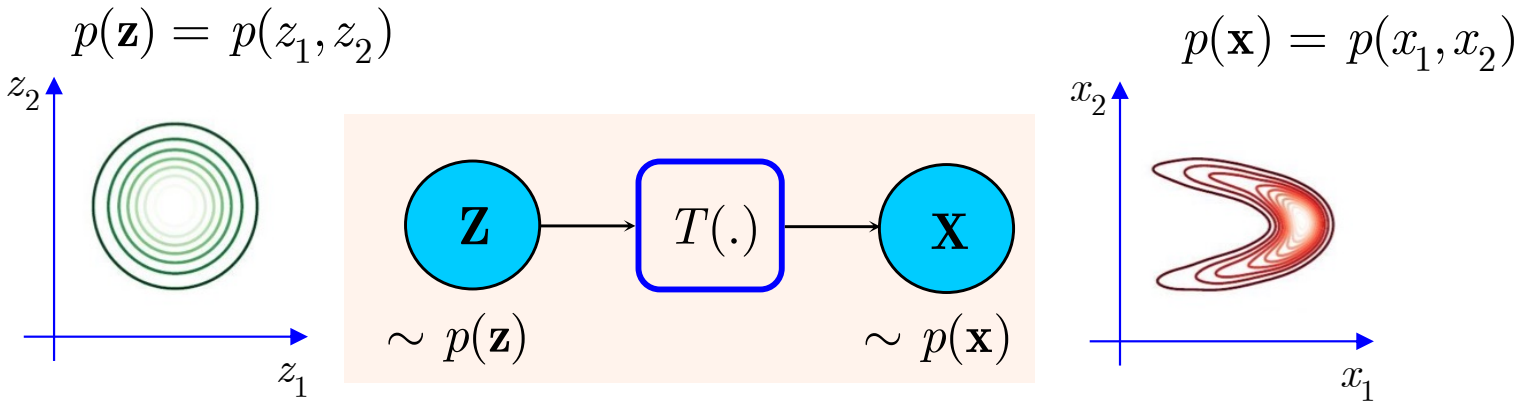
  - we draw MC samples from $p(x)$ as before
  - but instead of taking the mean of the function $\log r(x_p^{(i)})$ evaluated on these samples (which is unavailable for implicit distributions), we do so on a proxy function $\log r_\theta(x_p^{(i)})$ that is estimated through probabilistic classification as described above.

Scope

- Information theoretic measures for discrete/continuous variables
  - Entropy
  - Conditional entropy
  - Joint entropy
  - Relative entropy (KL-divergence)
  - Cross entropy
  - Mutual information
    - Additional topics:
      - f-divergence
      - practical computations from samples
        - Estimation of KL- and f-divergences by density ratio trick
        - Variational approximation of mutual information

Approximation approach: IT generative models

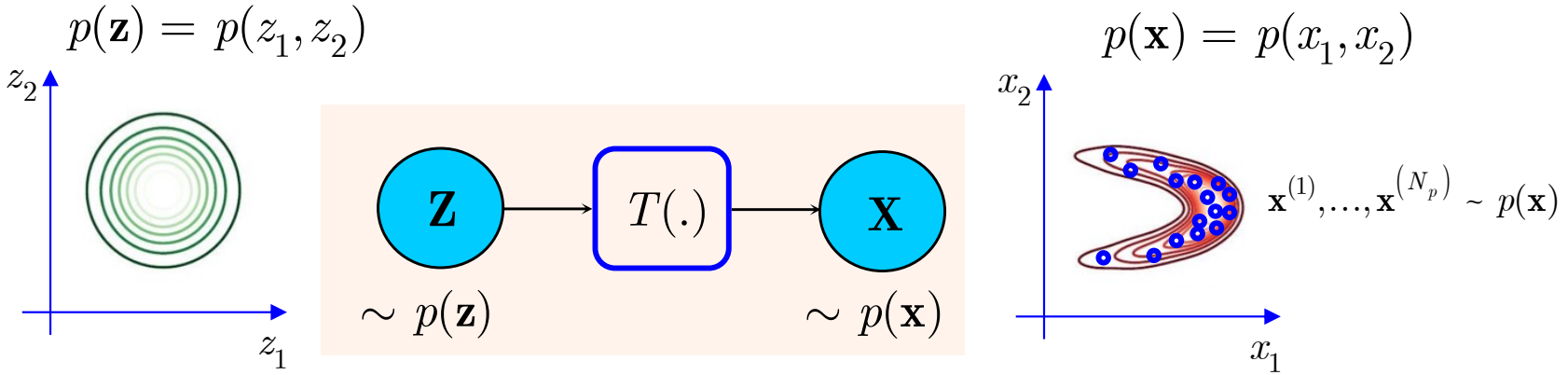
- Suppose we want to build a mapper (generator) that generates samples following the distribution $p(\mathbf{x})$ from the samples generated by the distribution $p(\mathbf{z})$



- If the distributions $p(\mathbf{z})$ and $p(\mathbf{x})$ were known analytically, we would easily compute a mapper (transform) $T(.)$ analytically (see Theme 1)

Approximation approach: IT generative models

- In practice: the distribution $p(\mathbf{z})$ is known and we can easily generate samples from it (no need to parametrize it) while the distribution $p(\mathbf{x})$ is only represented by samples $\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(N_p)} \sim p(\mathbf{x})$

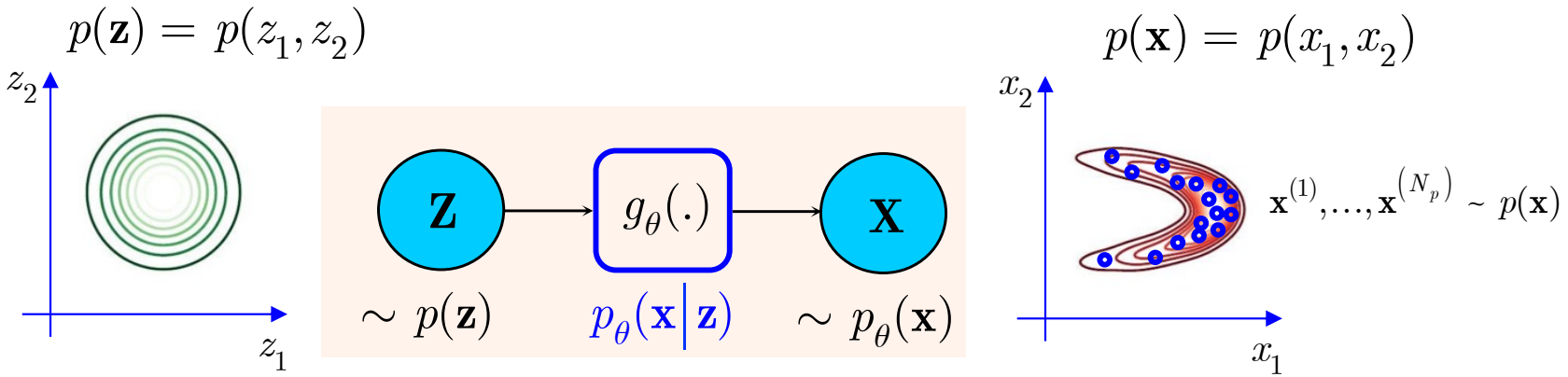


- Our challenge: to find a transformation $T(.)$ that would map the distribution $p(\mathbf{z})$ into the observed samples $\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(N_p)} \sim p(\mathbf{x})$

- The "empirical" distribution of observed samples is denoted as $p_D(\mathbf{x})$

Approximation approach: IT generative models

- Parametrization: we proceed by parametrizing an unknown transform $T(.)$ by some parametric function $g_\theta(.)$ with a set of parameters θ



- The distribution of generated samples (samples $\mathbf{Z} \sim p(\mathbf{z})$ passed via $p_\theta(\mathbf{x}|\mathbf{z})$) is

$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) d\mathbf{z}$$

- Our goal would be to find parameters θ, such that $p_\theta(\mathbf{x}) \to p_D(\mathbf{x})$

Approximation approach: IT generative models

- Possible strategies to find parameters θ such that: $p_\theta(\mathbf{x}) \to p_D(\mathbf{x})$
  - via KL-divergence minimization

$$\theta = \arg\min_\theta D_{\mathrm{KL}}(p_D(\mathbf{x}) \| p_\theta(\mathbf{x}))$$

We know how to do it and also properties of direct and reverse KLD

  - via maximization of mutual information

$$\theta = \arg\max_\theta I_\theta(\mathbf{X};\mathbf{Z})$$

$$I_\theta(\mathbf{X};\mathbf{Z}) = E_{p(\mathbf{x},\mathbf{z})}\left[\log \frac{p_\theta(\mathbf{x}|\mathbf{z})}{p_D(\mathbf{x})} \frac{p(\mathbf{z})}{p(\mathbf{z})} \frac{p_\theta(\mathbf{x})}{p_\theta(\mathbf{x})}\right]$$

$$= \underbrace{E_{p(\mathbf{x},\mathbf{z})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}_{\propto -\lambda E_{p(\mathbf{x},\mathbf{z})}\left[\|\mathbf{x}-g_\theta(\mathbf{z})\|_2^2\right]} - \underbrace{E_{p(\mathbf{x},\mathbf{z})}\left[\log \frac{p_D(\mathbf{x})}{p_\theta(\mathbf{x})}\right]}_{-D_{\mathrm{KL}}(p_D(\mathbf{x})\|p_\theta(\mathbf{x}))} - \underbrace{E_{p(\mathbf{x},\mathbf{z})}\left[\log p_\theta(\mathbf{x})\right]}_{H(p_D(\mathbf{x});p_\theta(\mathbf{x}))}$$

$$p(\mathbf{x},\mathbf{z}) = p_D(\mathbf{x})p(\mathbf{z})$$

Approximation approach: IT generative models

- Possible strategies to find parameters θ such that: $p_\theta(\mathbf{x}) \to p_D(\mathbf{x})$

$$\theta = \arg\max_\theta I_\theta(\mathbf{X};\mathbf{Z})$$
$$= \arg\max_\theta \left[-\lambda E_{p(\mathbf{x},\mathbf{z})}\left[\left\| \mathbf{x} - g_\theta(\mathbf{z}) \right\|_2^2 \right] - D_{\mathrm{KL}}\left(p_D(\mathbf{x}) \middle\| p_\theta(\mathbf{x}) \right) + H(p_D(\mathbf{x}); p_\theta(\mathbf{x})) \right]$$

- Since $H(p_D(\mathbf{x}); p_\theta(\mathbf{x})) \geq 0$

$$\theta = \arg\min_\theta \left[\lambda E_{p(\mathbf{x},\mathbf{z})}\left[\left\| \mathbf{x} - g_\theta(\mathbf{z}) \right\|_2^2 \right] + D_{\mathrm{KL}}\left(p_D(\mathbf{x}) \middle\| p_\theta(\mathbf{x}) \right) \right]$$

We reduced it to two metrics that we know how to compute

- Note a link to the direct minimization of KL- divergence